

Awash with cash, bereft of leadership

The National Institutes of Health has been granted impressive funding increases, but its lack of a permanent director is getting beyond a joke. The sooner the Bush administration plugs this gap with an inspired leader, the better.

One of the most momentous features of the federal budget proposal released this week by President George W. Bush (see pages 564–565) is its inclusion of funds that complete a doubling in funding for the National Institutes of Health (NIH), from some US\$13.6 billion in 1998 to US\$27.3 billion in 2003.

If, as expected, this request is endorsed by the US Congress, it will constitute a remarkable victory for the researchers, medical schools, patient advocacy groups and drug companies that have lobbied for this objective. It is unfortunate, however, that this immense build-up at the biomedical research agency is reaching its climax during a prolonged period in which it has had no permanent director.

The NIH director's office has become a politically sensitive spot, as issues surrounding embryo and fetal research embroil the agency in US abortion politics. That made it tough for President Bill Clinton to replace Harold Varmus, the former NIH director, when he vacated the post in December 1999. Given the political sensitivity of these issues, anyone taking the job would have been unsure whether their tenure would survive a change of occupant at the White House.

When Bush was elected, an early effort was made to fill the position, without success. One year on, the strain is now beginning to show. Ruth Kirschstein, the acting director of the NIH, has a unique mastery of the agency's workings — the product of a quarter-century of experience in its higher echelons. But although that has equipped her well to run a holding operation, Kirschstein would be the last person to suggest that she has the profile to fill Varmus's shoes on a permanent basis.

Any large institution requires inspired leadership. Awash with money it may be, but in a number of important respects the NIH is adrift. Major strategic projects, including the completion of the new

clinical research centre at the agency's Bethesda campus and the closing stages of the Human Genome Project, require the overview and input of a permanent NIH director. Varmus also worked hard to restore the scientific vitality of the NIH's intramural programme at Bethesda, which accounts for some 10% of the agency's budget, and attracted several top names to the campus. But some of these people have now left, without being replaced. The director of the NIH, had there been one, would also have played a major role in reassuring the public that the best science was being used to counter bioterrorism, in the wake of the anthrax attacks in the United States last autumn.

Perhaps the biggest problem, however, is that the directorship is vacant at a time when Tommy Thompson, head of the NIH's parent body, the Department of Health and Human Services, is working to expand his influence over the research agency's activities. It is understandable that Thompson, a former state governor, wants a say in running his department's crown jewel. But it seems to have escaped his attention that the NIH has attained this status precisely because it has avoided the politicization of its programmes and the bureaucratization of its operations that blight many US government agencies.

In his thinly disguised power grab for authority over such sundry matters as travel policy and press and congressional relations at NIH institutes, Thompson has turned what ought to be a time of great vision and opportunity at the NIH into a period of unseemly niggling and infighting. Various excuses have been quietly voiced for the failure of Thompson and the Bush administration to find a leader for the NIH. But with Bush's policy on embryonic stem cells in place and the administration into its second year, it is way past time for them to name an NIH director of stature, who will ensure that its incredible expansion is managed effectively and imaginatively. ■

A troublesome *ménage à trois*

France's largest research agency is in need of reform — but so are the universities that host its labs.

With 25,000 staff working in every area of science, the CNRS is the emblem of French research. So when the country's premier auditing body states that all is not well with the agency, those involved should sit up and listen.

The auditors' report charges that the CNRS — which gets 2.4 billion euros (US\$2.1 billion) of public funding annually — has had no strategic plan for the past six years (see page 563). On paper, this is undeniable, although to be fair to the CNRS, its lack of strategic direction owes as much to a repeated buffeting from successive politicians seeking to make their mark on the agency as to the CNRS's own failings.

The good news is that the agency's governing board this week endorsed a plan that correctly identifies the thorny issues — such as poor mobility and lack of autonomy for young scientists — that lie at the root of the CNRS's problems. The question

is whether this plan can be translated into action.

The CNRS has long been caught in an unhappy *ménage à trois* between the universities, which host its labs, and its paymaster, the research ministry. If French science is to be freed from this tug-of-love, each partner needs a better-defined role.

Addressing this problem should be a top priority for the new research minister who will be appointed following the parliamentary elections in June. With half of the country's scientists due to retire over the next decade or so, the minister will have an unprecedented opportunity to reshape France's research system.

Equally important to revitalizing the CNRS, and perhaps more challenging, will be the need to get the universities to address the twin themes of mobility and autonomy, and to lessen the teaching demands on top university scientists. Only then can the universities and CNRS labs begin to interact in a more productive way. ■





Budget bonanza
White House budget director lays out the bottom line
p564



Lines of enquiry
German parliament offers support for stem-cell research
p566



Divided issue
Studies fuel debate over use of mammograms
p567



Traveller checks
International Space Station lays down rules for tourists
p568

National audit slams CNRS over poor management and planning

Sally Goodman, Paris

France's largest and most prestigious research agency, the CNRS, has been severely criticized in a national audit for its alleged lack of strategic planning, weak financial management and flawed human-resources policy.

The audit says that the agency's cumbersome 800-member national committee, and the high degree of autonomy enjoyed by its 1,200 laboratories, prevent it from planning properly.

The criticisms are likely to lead to a major overhaul of the research agency and its governing committee — although the CNRS says that many of them are already being addressed in a new strategic plan that it will endorse this month.

The ability of researchers to move within the CNRS and to conduct interdisciplinary research is stifled, the audit says, by the inflexibility of the agency's discipline-based organization and evaluation system. The audit also charges that the CNRS — whose budget of 2.4 billion euros (US\$2.1 billion) makes it the largest supporter of basic scientific research in Europe — is not taking a leadership role in research involving other European countries.

The audit also criticizes the financial management of the agency, pointing out that its laboratories consistently underspend their allocated budgets, holding back the money for a rainy day. It adds that the agency faces a recruitment crisis, with half of its staff due to retire by 2020.

Despite recent reforms at the administrative level (see *Nature* 404, 426; 2000), the auditors charge that the CNRS and its parent research ministry have been "incapable of getting beyond the stage of collective reflection and group discussions" in enacting reform.

The current management team — Gérard Mégie, the agency's president, and Geneviève Berger, its director-general — were appointed in 2000 to lead a new management structure at the agency.

"This is a key moment for the CNRS," says Mégie. "We have done the thinking: now we need to translate this into action."

The agency's strategic plan says it will give



Under fire: Geneviève Berger (left) and Gérard Mégie say reforms at the CNRS are working.

priority to interdisciplinary work and to integration into the European Research Area. It introduces four-year contracts between CNRS management and its laboratories,

defining their budgets, objectives and evaluation procedures. Mégie says that he hopes the changes "will encourage researchers to take risks in their careers".

A 1998 plan by former research minister Claude Allègre to reduce the influence of the agency's national committee was abandoned in the face of fierce opposition from researchers (see *Nature* 396, 607; 1998). Berger says that the agency's own plan will reform the committee in a manner "no less radical than Allègre's, but better prepared".

The ministry of research, which itself is implicitly criticized by the auditors, says that many reforms have been implemented since the audit started two years ago. A new 'contract of objectives' between the ministry and the CNRS will be signed next month, defining long- and short-term objectives for the agency. "The contract will contain a number of performance indicators" says Ketty Schwartz, research director at the ministry. ■

Developing world gets patent aid

Declan Butler, Paris

An international organization is to be set up later this year to help health researchers in the developing world to navigate the complex maze of intellectual-property law.

Patent and licensing agreements often have a bad name in developing countries. They are widely associated, for example, with the lack of affordable access to drug treatment for diseases such as AIDS.

The new body, called the Management of Intellectual Property in Health R&D (MIHR), aims to assist governments in the developing world to negotiate better deals for drug access, as well as helping researchers there to protect their own ideas. The New York-based Rockefeller Foundation will provide start-up funding of US\$500,000 for the MIHR. With input from other donors, the MIHR's annual budget should reach US\$3 million by 2006.

The organization will employ a small staff of patent lawyers to give training and free legal advice to governments and researchers

in poor countries. It also plans to create free-access databases of patent information relevant to diseases such as malaria.

Ariel Pablos-Mendez, an associate director at the Rockefeller Foundation, says that the MIHR will try to persuade universities in rich nations to include provisions beneficial to poor countries in their licensing deals with drug companies.

John Kilama, president of the Global Biodiversity Institute, a Delaware-based organization providing information and training in biotechnology to developing countries, says that the MIHR can "fill a very important gap" by encouraging developing countries to make patent protection serve their own needs.

But Rosemary Wolson, an intellectual-property consultant at the University of Cape Town in South Africa, warns that it will not be easy for the new organization to win credibility, "bearing in mind how politically fraught the relevant issues are". ■

Bush goes to war as budget boosts R&D

Colin Macilwain, Washington

President George W. Bush announced his proposals for a 'war budget' on 4 February, which would see a record \$112 billion spent on research and development during the 2003 fiscal year, up 8% from this year.

The main components of the increase — which reflect the president's priorities of fighting the war against terrorism, homeland defence and economic growth — are a boost of \$5 billion for weapons development at the Pentagon, and an increase of \$3.7 billion in funding for the National Institutes of Health (NIH), announced last week (see *Nature* **415**, 459; 2002).

The Bush proposal will now be considered by Congress, which will set a final budget by October, when the 2003 fiscal year begins.

"The budget provides for an unprecedented level of support for research and development," says John Marburger, President Bush's science adviser. "It is the first time a president has requested more than \$100 billion for research and development."

The proposal would increase the US government's total spending on basic research by \$2 billion to \$25.5 billion — almost 60% of it at the NIH.

The scientific community was split along now-familiar lines in its initial reaction to this massive increase, with physical and environmental scientists feeling left out by the huge increases at the NIH.

Michael Lubell, head of public affairs at

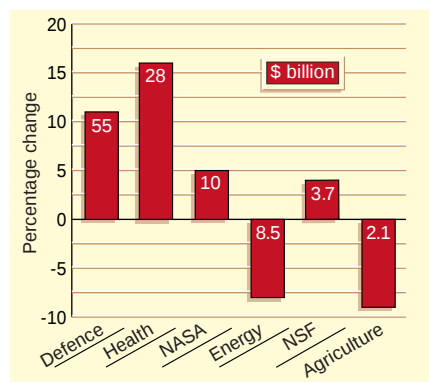
the American Physical Society, points out that when the NIH and the Department of Defense are discounted, spending on the rest of the research and development portfolio will fall. "At the National Science Foundation, for example, everything apart from mathematics and biology is actually cut," he says. "This isn't a budget to be thrilled about."

Sherwood Boehlert (Republican, New York), chairman of the Science Committee in the House of Representatives, said in a statement: "Research spending — excepting the NIH — would remain anaemic under this budget."

But Howard Garrison, a spokesman for the Federation of American Societies for Experimental Biology, says that the NIH increase is "what we've all been hoping for", and adds: "The big elements of this are all looking good."

The budget documents themselves put strong emphasis on plans, spearheaded by Mitch Daniels, the director of the White House Office of Management and Budget, to base budgets on a more thorough assessment of agency performance (see *Nature* **415**, 466–467; 2002).

The first part of this assessment, published in the budget, was harsh: the National Science Foundation (NSF) was the only agency in the entire government to get the 'thumbs up' for its financial management under the OMB's grading scheme. But NSF officials had to celebrate this accolade with imitation champagne from California, after



Battle stations: Bush's 2003 budget proposal takes the fight to health and defence.

their budget increase was held to a measly 3.4% (see opposite).

Marcus Peacock, associate director of the OMB, hints that the NSF will be rewarded for its management skills in future years. "This is the first year that we've attempted to link budget and performance," he says, predicting that additional responsibilities will be shifted to the agency soon.

Peacock also encourages scientists to drop their resistance to the OMB's plans to apply performance assessment to basic research. "I've been reading in a book that it took European mathematicians 300 years to accept the idea of negative numbers," he says. "I'm hoping we'll do better than that in getting people to accept performance assessment."

♦ www.whitehouse.gov/omb

Funding freeze leaves high-energy physics facing cuts

Geoff Brumfiel, Washington

The Department of Energy's Office of Science, which supports major research facilities across the country and is the United States' main sponsor of physics research, will receive no increase at all to its \$3.3-billion budget under President Bush's 2003 proposal.

Several programmes will be curtailed as a result of the funding freeze. In high-energy physics, for example, extra resources to upgrade the Tevatron at Fermilab, near Chicago, will come at the cost of closing down a fixed-target experiment at the Alternating Gradient Synchrotron at the Brookhaven National Laboratory in New York state.

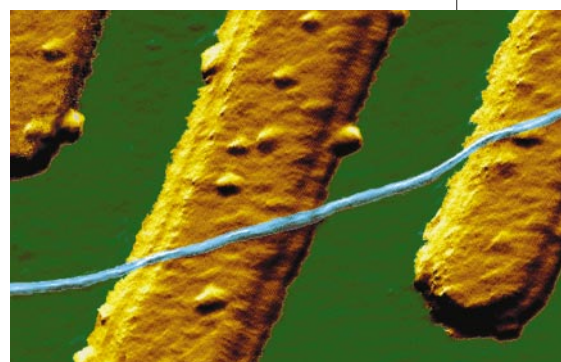
The proposed halting of this experiment has alarmed Brookhaven officials. "We had been promised running of the experiment at least through 2003, and now we wake up to find that they've just taken it away," says Thomas Kirk, an associate director at the

laboratory. "It's a very serious blow to our high-energy physics programme."

But John Marburger, the former Brookhaven director who now serves as Bush's science adviser, defends the emphasis on the Tevatron, which he says has a "window of opportunity" to search for the Higgs boson before Europe's Large Hadron Collider (LHC) begins operating in 2007. "Maybe they'll find it before the LHC switches on," he says.

Under the proposal, biological and environmental research at the energy department will decline sharply, from \$570 million this year to \$504 million next year, although department officials say that many of the cuts will affect projects that had been "earmarked" by Congress this year.

One bright spot in the Office of Science's proposal is \$45 million in new funding for four nanotechnology research centres, the first of which will be built at Oak Ridge National Laboratory in Tennessee.



Small wonder: nanotechnology is one of the few winners in the energy department's budget.

Although non-military science programmes at the Department of Energy will suffer, the National Nuclear Security Administration, which runs the department's nuclear-weapons research laboratories, will see its budget grow by 6% to more than \$8 billion.

NASA tunes in to nuclear power

Tony Reichhardt, Washington

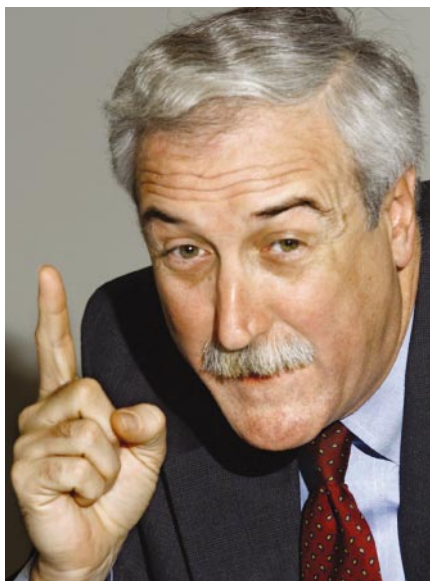
NASA plans to make major new investments in nuclear-powered rockets and spacecraft, its recently appointed head Sean O'Keefe said on 4 February as he unveiled the agency's budget proposal for the fiscal year 2003. But at \$15.1 billion, NASA's budget would be virtually unchanged from this year.

O'Keefe added that proposed missions to Jupiter's moon Europa and to Pluto will be scrapped to be replaced with a wider-ranging programme called New Frontiers, modelled after the agency's successful Discovery line of planetary spacecraft.

With a typical mission price tag of about \$650 million, New Frontiers will focus on studying the origins of life in the Solar System, as well as other priorities set by a forthcoming review of planetary exploration by the National Academy of Sciences. NASA will solicit mission concepts from laboratories inside and outside the agency this spring, and plans to choose the first winner next year.

The agency is set to make its first significant investments since the 1960s in space nuclear power (\$79 million) and nuclear rocket propulsion (\$46.5 million) during 2003. It sees these technologies as the best way to cut the time it takes to travel to the outer planets and to increase the working lives of probes and rovers sent to visit planets (see *Nature* 410, 626; 2001).

These and other technology investments account for much of the 12% increase in



Power up: Sean O'Keefe will invest in new ways to cut the travelling times of planetary probes.

NASA's space-science budget, to \$3.4 billion. According to the White House, the portion of NASA's total budget spent on research and development will grow by 5% to \$10 billion.

Funding for the troubled International Space Station drops by 13% to \$1.5 billion in the budget proposal. Agency managers have been given two years to solve its fiscal and managerial problems, or resign themselves to having a smaller, less capable station.

Extra duties offset NSF gains

Virginia Gewin, Washington

The National Science Foundation (NSF) has been put in charge of three new programmes in this week's budget proposal. But, despite being the only government agency praised for its financial management in a much-trumpeted recent White House assessment, it gets little financial reward.

The NSF — which funds most non-biomedical university research in the United States — receives an increase of 5% in the budget proposed by its director, Rita Colwell. Government officials acknowledge that this is really worth just 3.4% when its new responsibilities are taken into account.

Under the proposal, the NSF will take charge of the Sea Grant programme, transferred from the National Oceanic and Atmospheric Administration; the environmental education programme at the Environmental Protection Agency; and a new water-quality programme to replace

the toxic-substances hydrology programme at the US Geological Survey.

Apart from the new arrivals, the big winner at the NSF is the mathematics division (see *Nature* 414, 676; 2001), whose grant budget will grow by \$30 million to \$180 million. Other agency divisions, including astronomy, physics and chemistry, see their budgets fall by a few percentage points.

Two new major research projects that have been on hold for a year get the go-ahead in the budget. EarthScope, a network devoted to earthquake detection and research, will receive \$35 million and two prototype sites for the National Ecological Observatory Network will get \$12 million (see *Nature* 410, 854; 2001). Continued building of the Atacama Large Millimeter Array is allocated \$30 million.

And the impoverished graduate students supported by the foundation can look forward to an increase in their annual stipends, from \$21,500 to \$25,000.

Pentagon rise bypasses campuses

The budget will not do much for university researchers on the prowl for Pentagon funds — despite its proposed \$5-billion hike in research and development at the defence department.

All of the increase is earmarked for weapons-system development. The Department of Defense's \$1.4 billion in annual support for basic research will actually slip by \$10 million, while its applied-research funding slides by 7.5% to \$3.8 billion.

The main exception to this is the Defense Advanced Research Projects Agency, whose basic research programme will be boosted by almost a quarter to \$176 million, while its applied research budget grows by 15% to more than \$1.2 billion.

Peer-review programme is rewarded

The US Department of Agriculture said that it would double the funding next year for its National Research Initiative (NRI), to \$240 million.

Unlike most research spending by the agriculture department, NRI grants are competitively peer-reviewed, making the programme popular with plant geneticists and other university researchers.

The rest of the agriculture department's research and development programmes would be cut back from \$2.3 billion this year to \$2.1 billion. But the department proposes small increases in funding for research into bovine spongiform encephalopathy and foot-and-mouth disease.

Minority programmes win at NIH

As expected, Anthony Fauci's National Institute of Allergy and Infectious Diseases (NIAID) is the biggest winner from Bush's proposal to increase the budget of the National Institutes of Health (NIH) by 16%.

Under the proposal, the NIAID's budget would grow by 57% to \$4 billion, with most of the increase directed at bioterrorism research. The National Cancer Institute receives an increase of 12%, and the National Center on Minority Health and Health Disparities will receive 18%. Most other institutes would get an increase of 8–9%.

Funds for HIV/AIDS research at the NIH grow by 10% to \$2.8 billion, including a 24% boost in vaccine research and \$100 million for an international fund to fight HIV/AIDS, malaria and tuberculosis.

Technology scheme under siege

One of the Clinton administration's favourite research programmes — the Advanced Technology Program (ATP) at the National Institute of Standards and Technology (NIST) — will be almost halved in size under the Bush proposal.

The proposal cuts the programme's budget from \$185 million to \$108 million. A similar plan to shrink the ATP — which the administration regards as an unnecessary subsidy for work that industry should do itself — was blocked by Congress last year.

This time, the administration also proposes extra construction funding for NIST, providing \$67 million for new laboratories at its main campuses in Gaithersburg, Maryland, and in Boulder, Colorado.

German parliament backs stem-cell research

Quirin Schiermeier, Munich

Germany's stem-cell debate took a decisive turn on 30 January, when the parliament voted to let researchers use a limited number of human embryonic stem-cell lines.

The 340–265 vote got a cautious reception from researchers, some of whom are concerned about restrictions in the new law.

Germany will now permit embryonic stem-cell lines created before 30 January 2002 to be imported. Only research on projects that are ranked as a high priority by a new regulatory body will be allowed.

The regulator will ensure that permission is obtained from the parents of the embryo from which the cell line was extracted, and that there are no alternative means of doing the research. Applications will have to be vetted by a new national ethics committee.

The DFG, Germany's main research funding agency, announced last May that it was willing to fund research on stem cells. German law bans the creation (though not the import) of human embryonic stem cells for research purposes. But the DFG chose to delay funding until parliament had debated the issue (see *Nature* **411**, 875; 2001).

Despite the restrictions in the new law, its passage was welcomed by Oliver Brüstle and Otmar Wiestler, the University of Bonn neuroscientists whose application for DFG

funding helped spark the debate. Their grant to study the production of neural precursor cells from stem cells was approved the day after the parliamentary decision, but will remain frozen until the law comes into effect. Science minister Edelgard Buhlmann hopes that this will be before the summer.

"The ice is broken," says Brüstle. "We would have preferred a more liberal solution, but the new rules are helpful, provided they will not be used to create further delays." German researchers should have ready access to all the lines listed in the National Institutes of Health (NIH) registry, he says.

Günter Stock, head of research at Berlin-based pharmaceutical company Schering and a vocal supporter of stem-cell research, says, "It is the smallest possible yes, but I am grateful to parliament for the vote." Schering is not yet involved in stem-cell research, but Stock says it will consider cooperating with academic groups or start-up companies as soon as therapies are in sight.

The decision parallels that taken by US President George W. Bush last August, when he ruled that government funds could only be used to pay for research on stem-cell lines already in existence (see *Nature* **412**, 665; 2001). A subsequent survey by the NIH found that 72 lines held by 11 academic institutions and corporations met the criteria.

Wolfgang-Michael Franz, a cardiologist at the University of Munich who has applied for support for a project using stem cells, says that only about 10 of the lines on the NIH list are suitable for research — and that this number has not grown since last summer. ■



Eagle eye: the German parliament votes to allow stem-cell research, but under strict guidelines.

Drillers dig deep for microbes under the sea floor

Rex Dalton, San Diego

Microbes that live beneath the sea floor are the core issue for the latest expedition in the Ocean Drilling Program (ODP), which set sail from San Diego on 27 January.

The research ship will spend eight weeks drilling for sediment samples from the Pacific Ocean, mainly off the coast of Peru.

This is the first time that an ODP cruise has been dedicated to sedimentary biomass, says Steve D'Hondt, an oceanographer at the University of Rhode Island and co-director of the project.

Microbes living in seafloor sediment are thought to make up between one-tenth and one-third of all the living organisms on the planet, but very little is known about them. One of the expedition's goals is to find out how the microbes manage to survive in such an unpromising environment.

The international crew of 25 researchers and a similar number of technicians on board the *JOIDES Resolution* will recover sea-floor sediment and study the biomass it contains.

Better understanding of the microbes' metabolism should help to explain the

interaction among biomass, ocean chemistry and carbon sequestration — a natural process by which carbon is stored in the deep ocean, and which could possibly be tapped artificially to absorb greenhouse gases from the atmosphere.

The *Resolution* has been fitted with a new laboratory where researchers will initiate their study of the respiration of microbes found in the samples.

Previous, smaller-scale studies have found very little respiration — five to six orders of magnitude less than anticipated — from microbial life in the sediment, scientists say. "If one-third of life is in the sea floor, why is there so little respiration?" asks D'Hondt.

The expedition, costing about US\$8 million, is occurring as the ODP — which is supported by the US National Science Foundation and 21 partner agencies around the world — prepares to close down in 2003 after 18 years of operation. A new Integrated Ocean Drilling Program is to take its place, involving a broader international scientific and funding effort. ■

► www.oceandrilling.org



Big rig: researchers on the *Resolution* hope to find out more about elusive marine microbes.

Reviews spark debate over breast screening

Jonathan Knight, San Francisco

Faced with intense disagreement among epidemiologists over the best way to measure the health benefits of mammograms, the United States may be preparing to edge away from its long-standing policy of officially encouraging their use.

A review by the Copenhagen-based Nordic Cochrane Centre, published last October, questioned some of the studies cited to justify the use of mammography (see *Lancet* **358**, 1340–1342; 2001). In response to this, the US National Cancer Institute (NCI) now says it may revisit its current recommendation that all women over the age of 40 should receive regular mammograms.

And as public doubts grow over the robustness of the evidence on which the recommendation is based, two senators with strong influence over health policy, Barbara Mikulski (Democrat, Maryland) and Tom Harkin (Democrat, Iowa), have pledged to hold committee hearings on the topic in the next few weeks.

The dispute about the evidence centres on which data epidemiologists should measure when they study the outcome of cancer screening. If they measure cancer deaths only, the screening appears beneficial, some studies suggest. But if they measure overall mortality — including deaths that may result from intrusive cancer treatments, such as chemotherapy and surgery — the outcome becomes far murkier.

This week, researchers who believe that mammograms save lives hit back at the Cochrane review, arguing in the current issue of *The Lancet* (see *Lancet* **359**, 404–406; 2002) that its methods served to bury real, statistically significant evidence of health benefits from mammograms.

The authors of the new review, led by Olli Miettinen, an epidemiologist at McGill University in Montreal, argue that the Cochrane review misses this evidence by insisting that overall mortality is the only measure of its effectiveness. They argue that this approach obscures the small, but nonetheless real, benefits of the procedure. “To come to the conclusion that there is no benefit is wrong,” says team member Claudia Henschke, a radiologist at Weill Medical College in New York.

Most of the evidence that mammograms save lives by catching cancer early comes from studies conducted in the 1970s and 1980s involving almost half a million women in four countries. Five of the studies concluded that regular screening reduces breast cancer deaths by an average of 30%; two other studies found no benefit.

One reason for the difficulty in settling the disagreement is that the health effects being measured are so small, says William



Fresh assessments have cast doubt on data suggesting that regular mammograms save lives.

Black, a radiologist at Dartmouth Medical School in New Hampshire; even slight biases in determining the cause of death are likely to throw the results off. Black and his colleagues argue in a paper published this week that such biases are rampant in randomized cancer-screening trials (see *J. Natl Cancer Inst.* **94**, 167–173; 2002).

On 23 January, the NCI's Physician Data Query Screening and Prevention Editorial Board, an independent panel of experts set up to evaluate the literature on cancer screening for doctors and patients, agreed

that the Cochrane review had raised some valid questions. “The bottom line is uncertainty,” says panel member Donald Berry, a biostatistician at the University of Texas M. D. Anderson Cancer Center in Houston.

But the NCI will stick to its current screening recommendations, at least until the screening and prevention board's report is released in March, says NCI cancer-prevention chief Peter Greenwald. “I'm not sure whether we would consider revising the guidelines, but we will certainly consider the board's report,” he says. ■

Climate lobby group closes down

Virginia Gewin, Washington

The Global Climate Coalition (GCC), the main US lobby group opposed to mandatory cuts in greenhouse-gas emissions, has shut up shop.

In a statement on its website, the GCC says that, in effect, it is declaring victory and going home. With the Bush administration rejecting the Kyoto Protocol on climate change and instead backing a national strategy that relies on technology, rather than regulation, to cut greenhouse-gas emissions, the GCC claims to have achieved its main objectives.

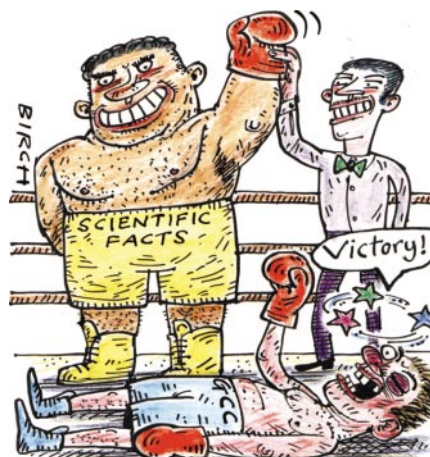
But critics of the GCC point out that it has lost many of its best-known members — including DuPont, Shell, Texaco, Ford and General Motors — in the past two years. Many of these corporations have publicly acknowledged the dangers of global warming, which the GCC played down for the 13 years of its existence.

“After a while, there weren't many leading companies left that wanted to be associated with the view the GCC was advancing,” says Lester Brown, founder and former chairman of the Washington-based Worldwatch

Institute, an environmental lobby group.

Eileen Claussen, president of the Pew Center on Global Climate Change, a Washington-based group that works with industrial companies that believe climate change to be a problem, says that the United States is still likely to take various actions to reduce emissions, and that the GCC's talk of victory is premature. ■

♦ www.globalclimate.org



Canada's innovation proposals build up head of steam

Toronto The Canada Foundation for Innovation, the agency set up in 1997 to help to revive the country's research infrastructure, has announced the 280 recipients of its latest round of awards, totalling Can\$779 million (US\$490 million).

The foundation plans to invest Can\$3.15 billion in university buildings and equipment by 2010, and last week's announcement takes it halfway to meeting that goal. The awards, announced on 30 January, totalled about twice as much as in previous rounds. "We had a very rigorous review process, but we also had really good proposals so we figured we should go for it," says David Strangway, president of the foundation. But Strangway warned that the spending spree will leave the foundation with less in the bank for future rounds.

♦ www.innovation.ca

Few want to be a millionaire as prize chances go begging

Munich Wunderkind scientists from all over the world appear to have missed the chance to apply for a generous funding award. The 29 winners of the German Sofja Kovalevskaja Prize received their awards last week, but the organizers admitted that only 109 researchers applied, giving each applicant a better than one in four chance of landing 1.2 million euros (US\$1 million) in funding.

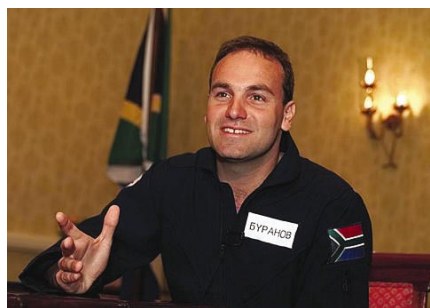
The one-off prize, named after a Russian mathematician, is designed to encourage foreign researchers to work in Germany (see *Nature* 409, 652; 2001). It is administered by the Alexander von Humboldt Foundation using a government fund generated by the sale of mobile-phone licences in 2000.

Many young scientists may have been unaware of the awards because of the limited time the government gave for the scheme's administration.

No alcoholics, please, we're in space

Washington Liars and habitual drinkers are not welcome on the International Space Station, say guidelines released last week.

The agreement between the US, European, Russian, Canadian and Japanese space agencies outlines conditions that must be met by space tourists, or any other visitors to the station who are not professional astronauts. Amateur astronauts, such as Mark Shuttleworth, the South African Internet millionaire who is due to visit the station this April, must be physically fit, able to cope under stress and have good interpersonal and communication skills.



Fit to fly: future 'space tourist' Mark Shuttleworth needs to cope under stress.

Disqualification criteria were also stated. Delinquency, criminal behaviour, drug addiction or habitual drunkenness will each be grounds for refusing entry to the station. Persons indulging in "notoriously disgraceful conduct" will also be barred, although there were no examples of who might fall into this category.

♦ www.nasa.gov/hqpaio/isscrewcriteria.pdf

Centre set to target genes behind depression

Munich Germany's Max Plank Society has signalled a new willingness to collaborate with industry by agreeing to host a new genotyping centre at its Munich-based institutes for psychiatry and biochemistry.

In an agreement signed on 4 February, GlaxoSmithKline committed to support the infrastructure and staff of the new Genetic Research Centre, and the Bavarian government pledged 4 million euros (US\$3.5 million) over the next four years for research.

The centre will generate clinical, genomic and functional data in an attempt to identify genes associated with complex disease. This work will include a study of 1,000 individuals with unipolar depression.

FBI goes electronic in anthrax inquiry

Washington The Federal Bureau of Investigation (FBI) has asked 30,000 microbiologists for help in its anthrax investigation. The bureau's plea came in an e-mail sent to the American Society for Microbiology (ASM). After consulting with its lawyers, the society forwarded the message to its entire US membership.

The message says that whoever mailed the anthrax spores to two news outlets and two senators last October is likely to be a "stand-offish" person with access to a laboratory.

"Whoever is responsible for the attacks has to have some scientific knowledge, and the ASM's members will have insights that might be able to help us," says Van Harp, assistant director of the bureau's Washington regional office. Harp would not say whether the FBI had received any responses.

♦ www.asmsa.org/pcsr/fbirequest.htm

Virologists struck down by virus at virology meeting

London Virology researchers attending a conference suffered a nasty experience after unwittingly taking their work home with them.

Some 40 delegates went down with vomiting, diarrhoea and fever in the days following last September's meeting of the European Society for Clinical Virology held in Lahti, Finland. The offender was quickly identified as a Norwalk-like virus, a common cause of such outbreaks. This was an example of a food- and water-borne infection — the very subject of the conference session at which the virus struck.

The exact source proved harder to pinpoint. "It was a buffet, and people were picking here and there," says Carl-Henrik von Bonsdorff, a virologist at the University of Helsinki. "But the most probable source was the salad." Von Bonsdorff reported the results of an investigation into the outbreak at another, so far incident-free, society meeting in London last month.

X marks the spot for Stanford geneticist

San Diego Bio-X, the interdisciplinary research facility being developed at Stanford University in California, has appointed developmental biologist Matthew Scott as its chairman.

Scott, who is already based at Stanford, is perhaps best known for his 1983 co-discovery of the homeobox, a family of genes that plays a crucial role in the development of the body plan of many animals. He started his five-year term at Bio-X on 1 January.

Bio-X aims to bring together engineers, biologists, medical researchers and physical scientists to develop new medical technologies. Scheduled for completion in June 2003, the centre was made possible largely by a US\$150-million pledge from Silicon Valley entrepreneur James Clark. An anonymous donor also gave \$60 million. But last year, Clark suspended \$60 million of his pledge as a protest against President George W. Bush's policies on stem-cell research, and one of Scott's challenges will be raise that money from elsewhere.

♦ cmgm.stanford.edu/biochem/biox



Model of the future: Stanford's Bio-X facility.

Planetary portraits

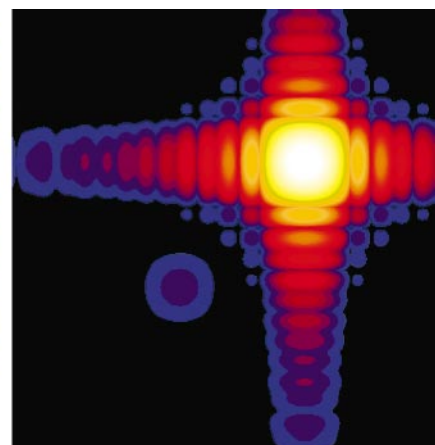
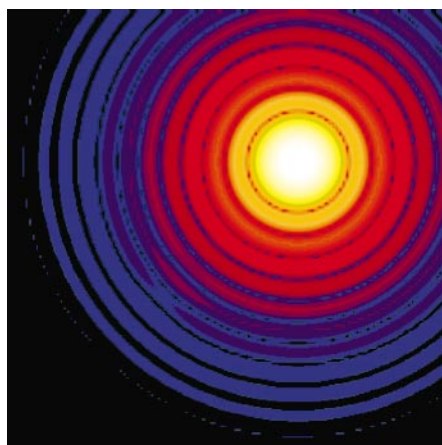
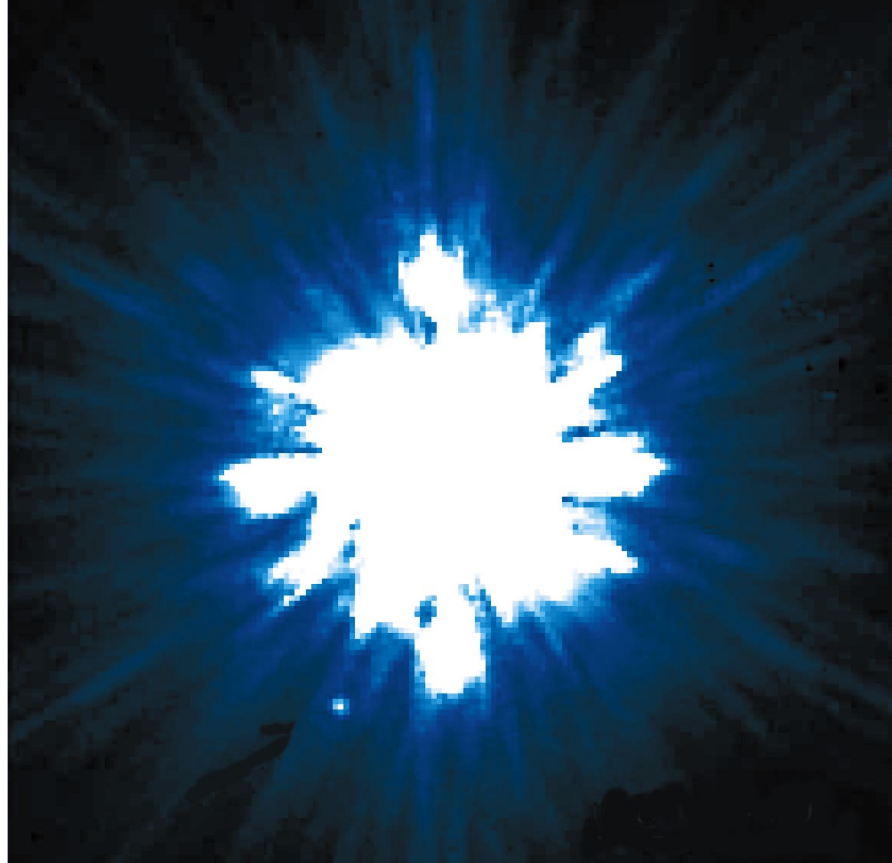
The list of known planets outside our Solar System grows by the week. Thanks to new technologies, a host of pictures of these worlds could soon be available. Tony Reichhardt investigates.

Soon after the first planet orbiting a Sun-like star was discovered, Dan Goldin, then administrator of NASA, issued a challenge. At a meeting of the American Astronomical Society (AAS) in San Antonio, Texas, in January 1996, he called on the assembled scientists to find a way, within 25 years, to take a picture of an Earth-sized planet orbiting a nearby star with a resolution good enough to see oceans, clouds, continents and mountain ranges. "And if that doesn't get your heart pumping," Goldin added, "I don't know what will."

Hearts may have pumped, but eyes also rolled. Surely Goldin understood the daunting technical problems. An Earth-sized planet orbiting a nearby star like our Sun would be lost in the glare from its parent, which would shine 10 billion times more brightly. Even a Jupiter-sized planet would be a billion times dimmer than its stellar companion.

But state-of-the-art space cameras can handle contrast levels of only around a million to one. Images of oceans and mountains were out of the question, and many astronomers doubted whether even crude imaging of extrasolar planets was feasible in the foreseeable future.

Fast forward six years to last month's annual AAS gathering, held in Washington. Extrasolar planet-hunting is now big business. Seventy-five have been discovered, and the list is growing by the week. Once relegated to back-room sessions, discussions of extrasolar planets filled the largest conference halls. And thanks to new concepts in telescope design, there was serious talk of moderately priced space missions that could return visible-light pictures of planets within five or six years. Goldin's pictures of alien



Brave new worlds: a brown dwarf (bottom left, top picture) next to a star. Simulations of new telescope apertures (bottom pictures) show how dimmer objects could be imaged. A square aperture (right) reveals a planet that is hidden by diffracted light when a circular aperture is used (left).

oceans remain mission impossible for now, but the first visible-light images of Jupiter-sized planets may be taken sooner than even optimists had thought.

Cutting the odds

Ground-based telescopes can already photograph brown dwarfs — objects that are intermediate in size between planets and stars — using infrared light, in which the brightness difference between the dwarf and the star it orbits is less pronounced. At last month's AAS meeting, Michael Liu of the University of Hawaii presented pictures (see above) of such an object orbiting 14 astronomical units from its star (1 AU is the distance from the Earth to the Sun).

But capturing visible-light images of Jupiter-sized planets will require new orbiting telescopes, as well as advances in imaging

technology. Such improvements are already in the offing, thanks to developments in coronagraphy — the technology of blocking unwanted light — and the techniques used to manufacture telescope mirrors.

The conventional approach to masking starlight is to place an opaque disk in the centre of the telescope's circular aperture, or to 'fog' selected areas of the telescope's mirror so that it does not reflect light from the star. But light that is diffracted by the edge of the aperture, along with light scattered by imperfections in the mirror, creeps in, creating a background glow that can obscure planets.

Rather than using a light-blocking disk, Peter Nisenson of the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts, and Costas Papaliolios of Harvard University, hope to solve the problem by using a square aperture (see



Giant hunters: Peter Nisenson (left) and Gary Melnick hope to image Jupiter-sized planets.

P. Nisenson & C. Papaliolios *Astrophys. J.* **548**, L201–L205; 2001). In contrast to the ring-like diffraction pattern created by circular apertures, their square aperture produces a cross-shaped pattern which can be rotated until the planet falls in one of the dark areas outside the bright cross (see left).

A modified version of this system is at the heart of the Extra-Solar Planet Imager (ESPI), a space telescope proposed to NASA by Gary Melnick, also at the Harvard-Smithsonian Center for Astrophysics. Launched from the space shuttle into Earth orbit, it would use a 1.5-metre-square aperture to take long exposures of 160 or more stars that are similar in age to our Sun and situated within about 50 light years of Earth. Melnick says the system could achieve the contrast necessary to image Jupiter-sized planets at 5 AU from their parent stars, which is Jupiter's distance from our Sun. For a handful of very close stars, pictures of planets that are only a few times more massive than Earth and only 1 AU from their stars might be possible.

But the benefits of the square aperture are not free. Fine-tuning the placement of diffracted light from the central star makes the whole system much more sensitive to imperfections in the optics. "Any stray light that can sneak into that dark region can fool you," says David Black of the Lunar and Planetary Institute in Houston, Texas, a veteran of the extrasolar-planet field.

So the ESPI project is counting on a new generation of mirrors, built with unprecedented smoothness and accuracy of shape, that have been developed for computer-chip manufacturers. In the quest to create ever-smaller microprocessors, the semiconductor industry has advanced the technology it uses to etch circuits onto wafers of silicon. Known as extreme-ultraviolet lithography, the new technique uses ultraviolet light to create circuit components that are less than 100 nanometres across. Optics companies have developed methods for making new and smoother mirrors to focus the beams of ultraviolet light. Melnick hopes that these techniques, which involve using ion beams to strip away layers of atoms one at a time,

will soon be applied to telescope mirrors.

So do advocates of the Jovian Planet Finder (JPF), another proposed planet-imaging mission. Principal investigator Mark Clampin of the Space Telescope Science Institute in Baltimore, Maryland, sees mirror imperfections as the biggest challenge. Rather than new apertures, he will use a series of super-smooth mirrors to search for Jupiter-sized planets around 50 stars, smaller Saturn-sized planets around 34 stars, and brown dwarfs around 2,000 stars. The JPF's planned vantage point is unorthodox: it would cling to the outside of the International Space Station.

Like the ESPI, the JPF has been proposed to NASA as a candidate for a Medium-Class Explorer (MIDEX) mission to be launched in either 2007 or 2008. NASA will select several MIDEX mission concepts for further study in April, and two winners will be selected by the end of the year. If chosen, the projects would receive a budget of up to \$198 million.

Computer corrections

A third mission concept, Eclipse, was presented at the AAS meeting by John Trauger of the Jet Propulsion Laboratory in Pasadena, California. Eclipse would use fairly conventional coronagraphy, and would not rely on the new super-smooth mirrors. Instead, it would use a computer to adjust the shape of a deformable mirror, and so continually correct for inaccuracies in the image caused by imperfections in the mirror's surface. Proposed to NASA's Discovery Program, which allows for missions costing up to \$300 million, Eclipse would be able to detect Jupiter-sized planets within 5 AU of bright nearby stars, and could even see Uranus-sized planets in a few systems.

If these missions appeal to NASA's review panels, they could serve as warm-ups to the Terrestrial Planet Finder (TPF), a proposed NASA mission planned for a 2014 launch and expected to cost between \$1.3 billion and \$2.8 billion. The TPF will be sensitive enough to image Earth-sized planets, and will also study their spectra, searching for signs of life such as the signature of molecular oxygen.

Four groups are currently in contention to build the TPF. The field will be narrowed to two this spring, with the design to be finalized in 2006 — all subject to the vagaries of NASA funding and the possibility of a merger with the European Space Agency's Darwin project, which has broadly similar aims.

Designs for the TPF and Darwin have focused on the use of nulling interferometers that operate in the infrared spectrum. These devices combine images from several telescopes. If the distances between them are chosen correctly, the starlight that reaches the different telescopes will be out of phase and will cancel out when the images are combined, leaving only the faint glow from a planet.

But working with infrared wavelengths has a drawback: a larger telescope aperture is required to achieve the same resolution that smaller apertures can obtain when working with visible light. So designs for the TPF and Darwin have called for an array of telescopes, either flying in formation or yoked to a boom-like structure as much as 100 metres long. Cryogenic cooling might also be required to cool the detector to prevent images being swamped by infrared radiation from the detector itself, a detail that further complicates the design. And some of the most interesting spectral signs of life lie in the visible, rather than infrared, spectrum.

With new ideas for imaging planets in visible light now emerging, opinions on the TPF have begun to change. Two of the four teams — those based at Ball Aerospace and Technologies in Boulder, Colorado, and Boeing-SVS in Albuquerque, New Mexico — are now considering TPFs that would operate in the visible spectrum. "The whole game is now a bit more open," says Black.

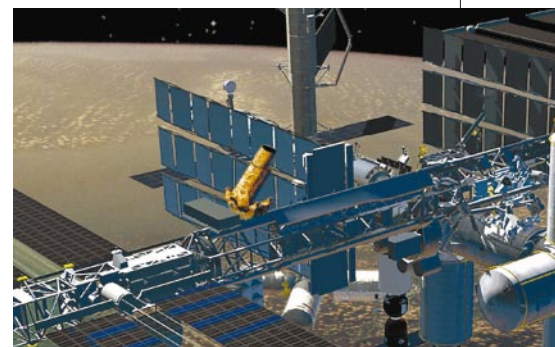
With a range of planet imagers on the drawing-board, and a revamped TPF under consideration, the next few years should be a busy time for planet-hunters. Goldin's extra-solar landscapes may yet be a few decades away, but his challenge is starting to look a little more realistic.

Tony Reichhardt is a contributing correspondent for Nature.

AAS Washington meeting abstracts ♦ www.aas.org

List of known extra-solar planets ♦ www.obspm.fr/planets

Terrestrial Planet Finder ♦ tpf.jpl.nasa.gov



Cling on: a mock-up of the Jovian Planet Finder attached to the International Space Station.

All at sea

The oceans are full of microorganisms, which are thought to cycle nutrients and mediate climate on a global scale. Despite these environmental consequences, marine microbial biodiversity remains poorly understood. Jon Copley reports.

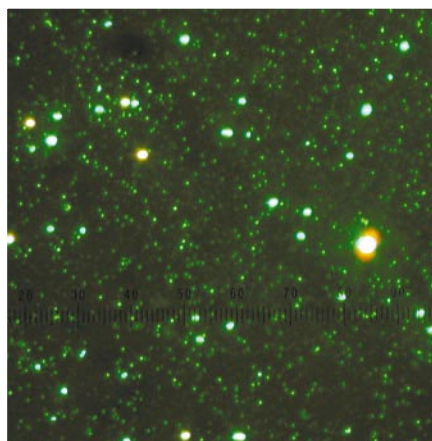
When it comes to mind-boggling numbers, marine microbiologists can give anyone a run for their money. The oceans are brimming with more than 3×10^{28} bacteria — that's about 100 million times more cells than there are stars in the visible Universe. But the real shock is that we have little idea what most of them are doing.

Only a tiny fraction of ocean-dwelling microorganisms have ever been cultured in the laboratory. Yet a better understanding of these microbes may be crucial to revealing how our planet works. "The role that the ocean plays in structuring the Earth's climate is largely driven by microorganisms," says Peter Burkill of the Plymouth Marine Laboratory in southwest England.

Around the world, researchers such as Burkill are now using a burgeoning range of techniques to sift through the myriad marine microorganisms. They hope to identify the key players and understand how these microbes influence the climate through their role in the cycles of elements such as carbon and nitrogen. Along the way, researchers are finding an unappreciated diversity of microbial life — and even new types of photosynthesis.

Microbial cells in the oceans are divided among three domains of life. There are the eukaryotes — single cells or simple colonies of cells whose nuclei are wrapped in membranes — such as green algae. Among the remainder — called prokaryotes — genetic and physiological differences define two further domains: the bacteria and the archaea. In addition, the oceans are teeming with viruses, most of which infect marine bacteria.

Archaea were first discovered in extreme environments such as the deep-sea vents that pour forth scalding sulphurous water from the ocean floor. They were thought to be restricted to such niches until researchers, including Jed Fuhrman of the University of



Southern California in Los Angeles^{1,2} and Ed DeLong of Monterey Bay Aquarium Research Institute in Moss Landing, California³, found them to be widespread in the ocean. "They make up typically 40% of deep-sea organisms — and the deep sea is by far the largest habitat on Earth," says Fuhrman.

Active life

Rather than being inert refugees expelled from deep-sea vents, most of these cells are metabolically active members of the plankton. Fuhrman has shown that the cells can mop up amino acids from vanishingly small concentrations in their surroundings⁴, and Markus Karner, then at the University of Hawaii in Honolulu, and his colleagues last year showed that the microbes contain plenty of ribosomal RNA — indicating that they are busy making proteins⁵.

Although the prevalence of active archaea may have been a revelation, even the partially familiar can still yield surprises at sea. One might think that by now marine biologists understood photosynthesis — the use of light energy to make sugars — in the oceans. But abundant new forms of photosynthesis and phototrophy — the generation of chemical energy from light — have recently been found among marine bacteria. Paul

Send in the marines: researchers such as Ed DeLong (pictured) are finding fresh traits among oceanic microbes (inset). Looking like stars, these organisms outnumber their stellar counterparts.

Falkowski of Rutgers University in New Brunswick, New Jersey, and his colleagues came across one of them by accident.

Falkowski's group was hunting for photosynthetic microbes called α -proteobacteria in the waters around deep-sea vents, reasoning that the incredibly faint glow from the vents might just be sufficient to support photosynthesis. The team used an infrared fast-repetition-rate fluorometer, which flashes pulses of light lasting fractions of a microsecond. These flashes cause chlorophyll and other photosynthetic pigments present to fluoresce. By measuring the wavelength and duration of this fluorescence, it is possible to deduce how much and what type of pigment is actively involved in photosynthesis.

No signs of photosynthesis were detected in water samples from deep-sea vents, but the team also tested their equipment on surface waters. Looking at the behaviour of the microbes in surface-water samples, the

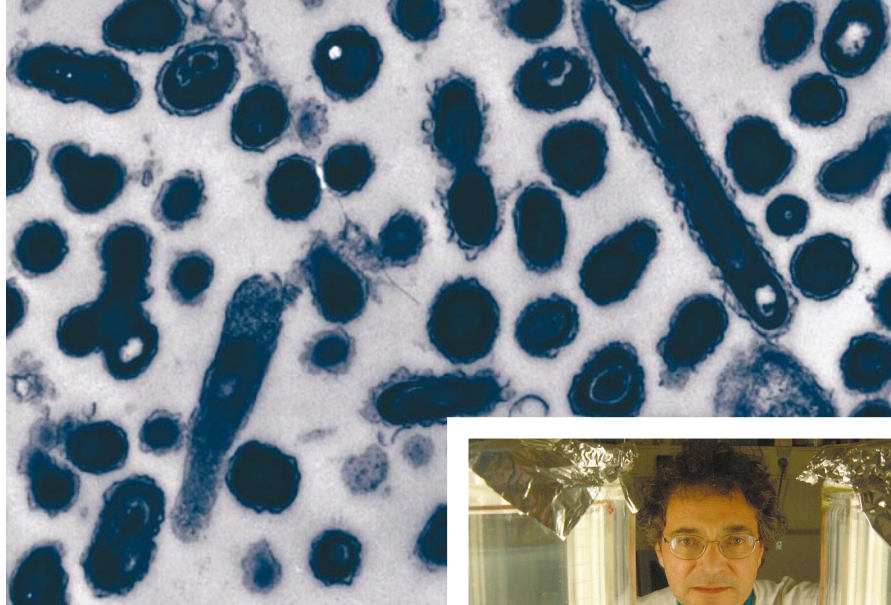
researchers estimated that α -proteobacteria account for up to 5% of the photosynthesis taking place⁶. "When we first saw these signals in the upper ocean, we thought there was something wrong with our machines," recalls Falkowski. But unlike traditional practitioners of photosynthesis, such as eukaryotic algae or cyanobacteria, the α -proteobacteria do not produce oxygen.



Jed Fuhrman spotted widespread archaea.

C. GOTTSCHALK, UCSB; INSET: J. FUHRMAN

L. HEWSON



Paul Falkowski has found unusual photosynthetic behaviour in some marine proteobacteria.

Instead they use oxygen to synthesize their chlorophyll and grow. This unusual variant of photosynthesis, called 'aerobic, anoxygenic', was previously thought only to occur among the α -proteobacteria in places such as beach sands and on the surfaces of seaweeds.

Falkowski and his colleagues have found the α -proteobacteria wherever they have looked in the surface waters of the Pacific and the Atlantic. "We have 30 strains of them in culture," says Falkowski. "We believe they comprise approximately 10% of all the bacteria in the ocean." The photosynthetic α -proteobacteria are relatively abundant in vast circular currents known as gyres, which tend to contain low levels of the dissolved organic matter that non-photosynthetic bacteria rely on as a source of energy.

Fixed return

Photosynthetic organisms usually 'fix' carbon by using light energy to turn inorganic carbon, often in the form of carbon dioxide, into sugars. Falkowski's α -proteobacteria also fix carbon, but in an unusual way. "They have a funny kind of metabolism called photoheterotrophy," he says. In other words, the microbes use light energy to help them break down organic matter, but they fix some inorganic carbon as well.

Carbon fixing is key to the way in which oceans absorb CO_2 — and so affect climate. But simply adding more carbon-fixing organisms to the ocean does not necessarily increase its capacity to absorb atmospheric CO_2 . This also depends on the processes that remove fixed carbon from surface waters. But the new bacteria add a fresh layer of complexity to the flow carbon takes through the ocean.

Much of the carbon fixed in the ocean ends up as dissolved organic matter, released when microorganisms die. This organic matter can be used by other microbes that may in turn be consumed by yet other microbes. In this way, the organic matter eventually finds its way to the tiny creatures known as zooplankton — a process known as the microbial



loop. The zooplankton subsequently die and some of their biomass sinks to the ocean floor, locking away the fixed carbon.

Some carbon is lost at each step in the microbial loop, as the organisms involved respire and release CO_2 . So knowing the length of the microbial loop is important for estimating the efficiency with which the ocean can process carbon.

The photosynthetic α -proteobacteria complicate matters further because they both use up dissolved organic matter in respiration, and can also add to it by fixing CO_2 . "They're actually making a loop within a loop," says Falkowski. This recycling of carbon keeps it in surface waters, rather than sinking to the ocean depths. That is good news for organisms that can use it, but — depending on the relative rates of the processes involved — potentially bad news for anyone hoping that the oceans will quickly put excess atmospheric CO_2 out of harm's way.

In this issue of *Nature*⁷, DeLong's group, working with researchers at The Institute for Genomic Research in Rockville, Maryland, show that the α -proteobacteria are far from the only marine microbes to practise aerobic, anoxygenic photosynthesis. After analysing bacterial gene sequences from seawater samples, they concluded that a multitude of distantly related bacteria can do the trick, including some unrelated to any known species. Other details of the physiology of these microbes remain unknown — clearly there is much work to be done before their effect on carbon cycling and other environmental processes is fully understood⁸.

In earlier work, DeLong and his colleagues came across another surprise when trawling through the genome of a member of a group called the γ -proteobacteria. They spotted a gene very similar to one that codes for the pigment rhodopsin in archaea⁹. Rhodopsin absorbs light, a reaction used by some archaea to drive their proton pumps — part of the molecular machinery used by cells to make chemical energy. But this type

of phototrophy had never before been seen in bacteria.

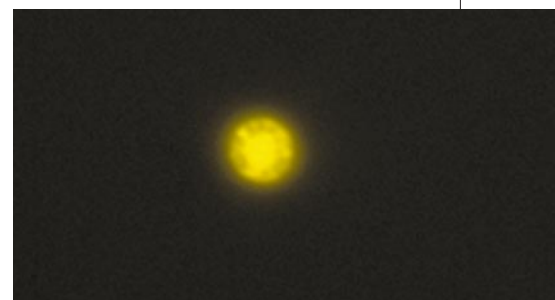
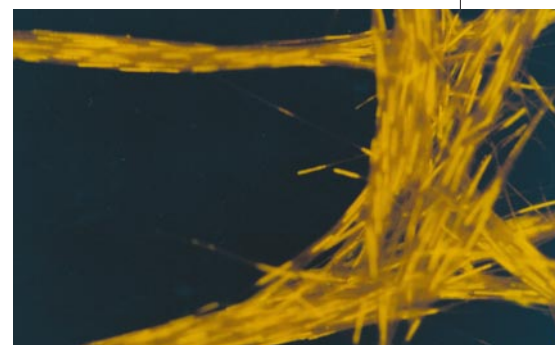
DeLong's group has since shown that this new phototrophic pigment, dubbed proteorhodopsin, is present in bacteria from the coastal waters of Monterey Bay, the central North Pacific and the Southern Ocean¹⁰. The researchers have also identified different versions of the pigment tuned to the different lighting conditions in the waters in which the bacteria involved are found. As with Falkowski's α -proteobacteria, this previously overlooked form of solar power may be widespread in the oceans — and significant for the flow of carbon through them.

In the loop

The jury is still out on whether the proteorhodopsin-containing bacteria actually fix carbon. But regardless of the verdict, the bacteria's ability to generate chemical energy using sunlight reduces their need to obtain energy by oxidizing dissolved organic matter. So these bacteria may also make their mark on the microbial loop.

Nitrogen is another vital element in the ocean, stimulating the microbial growth. "How much CO_2 can be removed by the oceans through photosynthesis is dependent on fertilization by nutrients such as nitrogen," says Jonathan Zehr of the University of California, Santa Cruz. Unfortunately, the most abundant form of nitrogen — dissolved N_2 gas — is useless to most marine life. It first has to be trapped in a biologically accessible form by nitrogen-fixing organisms.

The cyanobacterium *Trichodesmium*, which sometimes forms filamentous colonies just visible to the naked eye, has long been known to use a nitrogenase enzyme to



Gas traps: filamentous *Trichodesmium* (top) and single-celled cyanobacteria both fix nitrogen.

convert N_2 into ammonia. But it is not abundant enough to account for the amount of nitrogen being fixed in the ocean, suggesting that there are other nitrogen-fixers out there.

Zehr and his colleagues have found messenger RNA coding for part of the nitrogenase enzyme in several strains of smaller unicellular cyanobacteria 3–10 micrometres in diameter, and have shown that these microbes can fix radiolabelled N_2 (refs 11,12). Once again, these newly appreciated microbes seem to be widespread in surface waters¹³. Their role in nitrogen fixation could be equal to that of *Trichodesmium*, Zehr suggests.

Zehr is convinced that yet more nitrogen-fixers are waiting to be described. “We’re finding other culprits,” he says. Falkowski has a hunch that his photosynthetic α -proteobacteria could be involved. And Zehr has just received funds from the US National Science Foundation to examine symbiotic relationships in which the new nitrogen-fixing cyanobacteria live inside larger single-celled algae. “These may turn out to be the next really exciting story,” he says.

Variety show

Beyond the bacteria and archaea, molecular techniques are now revealing a hitherto unrecognized diversity of small eukaryotes in the picoplankton — cells less than 3 micrometres across. By analysing ribosomal RNA gene sequences, researchers led by David Moreira of the Miguel Hernández University in Alicante, Spain, have unearthed a host of new eukaryotes in the deep sea of the Southern Ocean^{14,15}. “I have a feeling that we will realize that their diversity is comparable to that of the prokaryotes,” predicts Moreira, who recently moved to the Pierre and Marie Curie University in Paris.

Using the same approach, researchers led by Daniel Vaultot of the Roscoff Biological Station in Brittany, France, have found a similarly diverse assemblage in the surface waters of the Pacific¹⁶. The eukaryotes are important grazers of other picoplankton, says Vaultot.

The picoplankton also contains photosynthetic eukaryotes, whose taxonomy and physiology is poorly known — although a European programme called PICODIV, coordinated by Vaultot, may soon start revealing their secrets.

At the very bottom of the size spectrum are the marine viruses. “In an average seawater sample you’ll get around a million bacteria per millilitre, but you’ve usually got about 10 million viruses,” says Willie Wilson of the UK Marine Biological Association in Plymouth. Most infect bacteria, preventing any one type from dominating the plankton — if the population of a particular bacterium



Wipe out: tiny marine viruses (arrowed, inset) kill the alga *Emiliana huxleyi*. They help to control its milky blooms (main picture), and in doing so influence cloud formation.

explodes, it becomes a target for a viral epidemic that culls its numbers.

Viruses can also transfer genes between different organisms. “That kind of genetic transfer has the potential to have a profound effect, even if it’s a one-in-a-billion event,” says Fuhrman. “It’s a subject that is wide open for discovery.” One intriguing possibility is that the proteorhodopsin gene found in DeLong’s γ -proteobacteria may have been transferred from archaea in this way.

Bloom and bust

The viruses also attack single-celled algae. In unpublished work, Wilson has isolated one virus that pervades the huge blooms of *Emiliana huxleyi* that can turn patches of the North Atlantic milky white. “We have 10 different strains of the virus from coastal waters in the English Channel,” he says, adding that the viruses are one factor causing the eventual death of the algae in these blooms.

When infected cells die, they release dissolved organic matter. So viruses might influence the microbial loop and the carbon cycle. “There’s no question that they play a part in the big picture,” says Fuhrman.

Wilson has also shown that when viruses kill off algal blooms, large quantities of dimethylsulphide (DMS) gas is released into the atmosphere. DMS stimulates cloud formation, increasing the proportion of solar radiation reflected back into space — which can have a dramatic cooling effect on climate.

In further unpublished studies, Burkill and his colleagues have found that some α -proteobacteria are also involved in this process, converting dimethylsulphoniopropionate (DMSP) released by dying algae into the cloud-forming DMS. “We’ve identified

the main route by which DMS is produced in sea water,” Burkill claims. His team used flow cytometry, originally developed to sort and study human cancer cells, to pinpoint the DMSP-converting microorganisms.

Findings such as this illustrate the gains being realized now that a diverse set of biological tools is being applied to the study of marine microbial biodiversity and the roles that the microbes perform in ocean ecosystems. “What’s fantastic is that the community has been able to bring to bear such a range of techniques,” says Burkill. Even so, marine biologists are still barely scratching the surface of understanding the microbial processes that determine the cycling of carbon and nitrogen in the oceans — information that will be crucial to determining the influence that marine microorganisms exert over our climate.

Jon Copley is a freelance writer in Southampton, UK.

1. Fuhrman, J. A., McCallum, K. & Davis, A. A. *Nature* **356**, 148–149 (1992).
2. Fuhrman, J. A. & Davis, A. A. *Mar. Ecol. Prog. Ser.* **150**, 275–285 (1997).
3. DeLong, E. F. *Proc. Natl Acad. Sci. USA* **89**, 5685–5689 (1992).
4. Ouverney, C. C. & Fuhrman, J. A. *Appl. Environ. Microbiol.* **66**, 4829–4833 (2000).
5. Karner, M. B., DeLong, E. F. & Karl, D. M. *Nature* **409**, 507–510 (2001).
6. Kolber, Z. S., Van Dover, C. L., Niederman, R. A. & Falkowski, P. G. *Nature* **407**, 177–179 (2000).
7. Béjà, O. et al. *Nature* **415**, 630–633 (2002).
8. Karl, D. M. *Nature* **415**, 590–591 (2002).
9. Béjà, O. et al. *Science* **289**, 1902–1906 (2000).
10. Béjà, O., Spudich, E. N., Spudich, J. L., Leclerc, M. & DeLong, E. F. *Nature* **411**, 786–789 (2001).
11. Zehr, J. P., Mellon, M. T. & Zani, S. *Appl. Environ. Microbiol.* **64**, 3444–3450 (1998).
12. Zehr, J. P. et al. *Nature* **412**, 635–638 (2001).
13. Neveux, J., Lantoin, F., Vaultot, D., Marie, D. & Blanchot, J. *J. Geophys. Res.* **104**, 3311–3321 (1999).
14. López-García, P., Rodríguez-Valera, F., Pedrós-Alió, C. & Moreira, D. *Nature* **409**, 603–607 (2001).
15. Moreira, D. & López-García, P. *Trends Microbiol.* **10**, 31–38 (2002).
16. Moon-van der Staay, S. Y., De Wachter, R. & Vaultot, D. *Nature* **409**, 607–610 (2001).

The 15% solution for majority health concerns

Medical journals should devote more space to issues that affect the developing world.

Sir— Amid much fanfare, the World Health Organization (WHO) announced last summer that it had convinced a consortium of medical publishers to provide physicians and scientists in the developing world with cheap access to research journals, via the Internet^{1,2}. This plan was characterized as “a real breakthrough” by WHO director-general Gro Harlem Brundtland, on 9 July 2001 (see www.who.int/director-general/speeches/2001).

But how much of a breakthrough is it? When the eager doctors and investigators finally open their new journals, they will find the articles inside almost exclusively geared to the medical problems of patients in developed countries, often discussing medications and technologies not available in underdeveloped regions of the world.

For example, a recent issue of the *New England Journal of Medicine* (a publication also distributed free over the Internet to low-income countries, though not a member of the new consortium) presented a clinical trial demonstrating that HIV patients responding to antiviral therapy — consisting of at least three powerful drugs — may no longer require precautionary antibiotics to guard against pneumonia³. The catch is that in most of the world, patients with HIV cannot afford even one of these medications. As an editorial accompanying this article observed, the results of the study “are good news for people living with AIDS but they also make the gulf in treatment between rich and poor countries that much more glaring and unacceptable”⁴.

If journal editors really want to make an enduring contribution to the developing world, they should focus on the content of their publications, not just the distribution. A good first step would be to commit a consistent number of pages — say 15% — to articles addressing the medical needs and concerns of underdeveloped regions. By acquiring a more global perspective, medical journals would accomplish several important benefits.

First, and most important, the journals would provide articles of interest and relevance to doctors, scientists and patients in these regions of the world, creating a body of data that could guide practice in the same way that clinical research has improved the treatment of patients in developed countries.

Second, articles focused on health care in underdeveloped regions would enable journal readers to become better acquainted with the unique needs and problems of medical care in these areas.

Less than 10% of the world's health-care expenditure is said to be devoted to illnesses that account for 90% of the global burden of disease, and it is likely that published research articles follow a similar pattern⁵. Increased focus on the health-care needs of poorer countries might encourage western physicians and scientists to think about problems they might not ordinarily encounter, and to generate fresh insights or novel approaches.

Finally, the new articles might help to inform the development of ethics guidelines for clinical research done by North American and European scientists in underdeveloped regions. These include thorny questions, such as whether it is ethical to conduct a study in which the control group of patients receives no medicine (the local standard of care) rather than an expensive medicine (the best care available). Greater familiarity with the needs and the resources of patients in poorer nations might help ethics committees as they wrestle with these difficult but vitally important decisions.

If journals were to commit 15% of their pages to medicine in the developing world, would their current readers be interested? Surprisingly, the answer may be a resounding yes. Western medical schools are reporting a groundswell of interest in international medicine; for example, at Harvard, a record number of medical

students — more than 50 — participated this past year in international health experiences, in countries such as Costa Rica, Bolivia, Vietnam and Zambia. The Association of Schools of Public Health, based in Washington DC, reports that the number of public-health students specializing in international health is now at an all-time high⁶. Meanwhile, organizations focusing on international health are receiving unprecedented recognition, such as Médecins sans Frontières, recipient of the 1999 Nobel peace prize.

By dedicating 15% of their pages to the medical concerns of the developing world, journal editors would not only show their genuine commitment to global health, but might also discover a Western readership eager for a broadened perspective, and an international medical community grateful for the opportunity to provide it.

David A. Shaywitz, Dennis A. Ausiello
Department of Medicine, GRB-740, Massachusetts General Hospital, 55 Fruit Street, Boston, Massachusetts 02114, USA, and Harvard Medical School, 25 Shattuck Street, Boston, Massachusetts 02115, USA

1. Malakoff, D. *Science* **293**, 189–190 (2001).
2. *Nature* **412**, 110 (2001).
3. de Quiron, J. N. *Engl. J. Med.* **344**, 159–167 (2001).
4. Girard, P. M. N. *Engl. J. Med.* **344**, 222–223 (2001).
5. Global Forum for Health Research. *The 10/90 Report on Health Research 2000* (<http://www.globalforumhealth.org/>).
6. Helsing, K. *2000 Annual Data Report: Applications, New Enrollments and Students, Fall 2000*. 41 (Association of Schools of Public Health, Washington DC, 2001).

World hasn't changed for the dispossessed

Sir— Understandably, following the outrages of 11 September 2001, the main scientific periodicals have covered its impact on aspects of science and the response of scientists to the fears that it generated. The annual News reviews in *Nature* (*Nature* **414**, 836–841; 2001) and *Science* both headline its aftermath, in terms of the impact on security and the economic downturn that the attacks have helped to accelerate. Their central theme is that the world changed on that day. It did not.

For two thirds of the world's population, ‘business as usual’ involves an ever-widening gap between hope for the future and expectation of any relief it will bring from poverty, disease and disaster. Fear of falling victim to natural calamities — and those generated by the lifestyles of a highly privileged minority — remains as strong as ever. Despicable as the perpetrators and those who motivated their actions were,

the attacks arose from the growing powerlessness of hundreds of millions of dispossessed people. Global communications ensure that they are confronted daily by what they lack, leading to a deep sense of unfairness and victimhood.

Scientists, whose work is enmeshed with emergence of the possible, should dwell on how they might help close that growing human fault line, as you discuss in your Opinion article “Timely messages for the South” (*Nature* **415**, 1; 2002), rather than raging at or cringing before the monstrosity that they have helped to nurture. Assisting the dispossessed to secure safe, dependable water supplies; to improve their agricultural yields; to rid themselves of endemic disease; to gain access to cheap energy and transport; and above all to acquire knowledge and the ability to solve their own problems is not a problem of cosmological or genomic proportions. It is a simple, human duty.

Steve Drury

Department of Earth Sciences, The Open University, Milton Keynes MK7 6AA, UK

Diagnostic testing fails the test

The pitfalls of patents are illustrated by the case of haemochromatosis.

Jon F. Merz, Antigone G. Kriss,
Debra G. B. Leonard and
Mildred K. Cho

Questions about the effects of patents and licensing are becoming critical in the United States, Europe and other developed countries as more genes are discovered and patented, and as genetic testing becomes an integral part of standard medical care. The award of patents for the diagnostic test for haemochromatosis, a progressive iron-overload disease, joins an ever-growing list of such tests that have been, or will very soon be, patented. We have found that US laboratories have refrained from offering clinical-testing services for haemochromatosis because of the patents. A lot of clinical study is needed to validate and extend the early discovery of a disease gene such as that for haemochromatosis, so our results give us reason to fear that limiting clinical testing will inhibit further discovery as well as the understanding that emerges naturally from broad medical adoption¹.

As highlighted by the looming patent battle over testing for a breast-cancer mutation between Myriad Genetics and the French Curie and Gustave Roussy institutes (see ref. 2), restrictive licensing and monopolization of clinical-testing services will not be limited to the United States for long. Indeed, four patents relating to haemochromatosis testing are pending in the European Patent Office, suggesting that the situation in the United States described here may soon spread to Europe.

Patent concerns

New human genes are being patented as rapidly as they are discovered^{3,4}. Gene patents generally cover the clinical diagnosis of mutations, as well as using the gene sequence in potential therapies. Setting aside the debate about the ethics of allowing any patenting of human gene sequences, many are concerned about the ramifications of gene patents for biomedical research and clinical medicine.

Unfortunately, there are few empirical data about the effects of patents on the translation of genomic discoveries into medical advances, so it is not clear how justified these concerns might be. Here, we present the results of a survey of US laboratories' adoption and use of genetic testing for hereditary haemochromatosis.

Hereditary haemochromatosis is a common autosomal recessive disease, affecting 1 in 200 to 1 in 300 people of northern European descent, with a carrier frequency of up



Medieval medicine. Haemochromatosis, thought of as a disease of the twenty-first century, can be treated by regular phlebotomy, or bleeding, a medical technique used back in the Dark Ages.

to 1 in 10 (ref. 5). As much as 80–85% of haemochromatosis is caused by the two most common mutant alleles of the HFE gene (C282Y and H63D). Haemochromatosis can be treated by periodic therapeutic phlebotomy (such as regular blood donation), so it is a candidate for population screening — and there is a potentially large market for clinical genetic-testing services.

We have discovered that many US laboratories began genetic testing for haemochromatosis before the patents were awarded, but

30% of those in our survey reported discontinuing or not developing genetic testing in the light of the exclusive licence granted on the patents covering clinical-testing services. This result raises obvious concerns about test quality, patient access to testing services, the costs of clinical testing, innovation of testing methods, and the potential for placing limitations on clinical research.

The US patents (numbers 5,712,098; 5,753,438; and 5,705,343) covering the HFE genetic test were first issued to Mercator

Genetics in early 1998. The patents grant the right to exclude others from testing for two mutations, C282Y and H63D. Mercator went out of business after spending about US\$10 million developing its patented method of positional cloning and discovering the association between HFE mutations and haemochromatosis. Progenitor merged with Mercator and was assigned its pending and issued patents. Progenitor then licensed the patents exclusively for clinical testing to SmithKline Beecham Clinical Laboratories (SBCL) for an up-front payment and guaranteed continuing fees worth around \$3 million. SBCL's exclusivity and payment guarantees continued until a kit became available for use by clinical laboratories. (Exclusive licensing of gene patents is common, particularly for clinical diagnostic uses⁶.)

Controlling interest

In the summer of 1998, SBCL began enforcing its patent rights. The company wrote to laboratories, stating its willingness to grant sublicences for an up-front fee of \$25,000 from academic laboratories, and 5 to 10 times more than this from commercial laboratories, plus royalties of as much as \$20 per test. In January 1999, SBCL was sold to Quest Diagnostics, but the sale was not completed until autumn 1999. During and after the sale, SBCL and Quest curtailed active enforcement of the patents, creating uncertainty for laboratories that were — or were interested in — performing HFE testing.

In April 1999, Bio-Rad Laboratories acquired the portfolio of pending and issued patents covering HFE and its mutations from Progenitor, subject to the exclusive clinical-testing licence held by SBCL. After the acquisition of SBCL in late 1999, Quest did not enforce the clinical-testing licence and, in October 2000, transferred it to Bio-Rad for terms that were not made public. In 2001, Bio-Rad began offering a test kit consisting of analyte-specific reagents for the C282Y and H63D alleles. According to several laboratory directors to whom we have spoken recently, Bio-Rad is now offering to license laboratories to perform testing without its kits — but at a cost that makes its kit more economically attractive than the laboratories' own tests, with up-front payments inversely proportional

Our data show that patents inhibited development and validation of clinical assays.

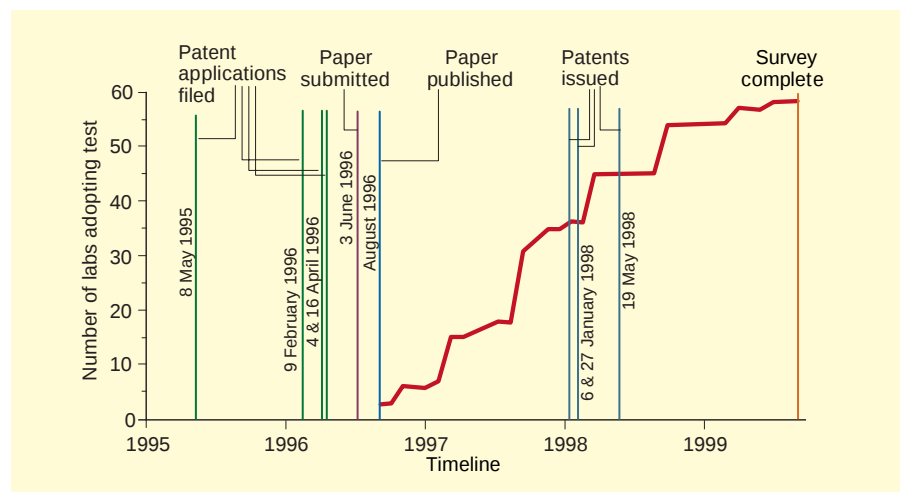


Figure 1 Timeline showing HFE patent applications, publication, patent grants and the cumulative number of laboratories at the times they began offering the clinical test following publication of the gene discovery in ref. 8.

to the testing volume of the laboratory, plus a per test fee of about \$20.

US laboratory survey

To understand how gene patents affected laboratories, we ran a pilot survey in November 1998 of laboratory directors and staff attending two national meetings⁷. Drawing on these results, we developed a comprehensive telephone survey, the results of which we report here, to assess the impact of the HFE patents and the SBCL licensing strategy on clinical-laboratory practices. We identified 117 laboratories in the GeneTests database (see www.genetests.org) and the Association for Molecular Pathology test directory that were or seemed capable of offering the HFE test. With approval of the University of Pennsylvania committee on studies involving human beings, one of us (A.G.K.) carried out interviews in July and August 1999. Snowball sampling (asking respondents for referrals) yielded 11 additional laboratories, for a total sample of 128. We completed 119 interviews (93%) — staff at 9 laboratories declined to participate.

Ninety-two respondents (77%) were laboratory directors, 12 were supervisors (10%), and the rest were other types of laboratory staff. Two-thirds (80) of the laboratories in our sample were affiliated with universities, hospitals or other non-profit institutions. Most (111; 93%) respondents reported knowing about the HFE patents; 61 had first heard about them from colleagues or at meetings, and 35 from SBCL's letter. Overall, 54 respondents received the letter, and the 58 laboratories performing HFE testing were more likely than those not offering the test to have received the letter (odds ratio = 4.4, $P < 0.001$).

Significantly, in September 1999, 31 (26%) laboratories reported that they had not developed and were not performing the test, and another 5 (4%) said they had

stopped performing the test. There was no difference between commercial and non-profit laboratories. Of these 36 laboratories, 22 reported that the patents were "the reason" and 10 said the patents were one of several reasons they had not developed or had stopped offering the test.

We believe that testing volume is a dominant factor in a laboratory's decision to carry out a particular clinical test. For hospital-based laboratories in the United States, the non-reimbursed expenses incurred by sending samples out to other laboratories for testing can be very high, motivating institutions to develop in-house tests if testing volume justifies the costs of development. Although we did not ask for specific reasons why laboratories were not performing HFE testing, it is likely that low test volume was one of the factors.

In sum, the patents on HFE had a measurable effect on the development and availability of HFE testing services in the United States, as many laboratories that had the capability to perform the test reported not doing so because of the patents.

We asked respondents when they started offering the HFE test, and present a timeline (Fig. 1) that covers the filing of patent applications, publication of the HFE discovery⁸, and issuance of the patents. Superimposed on this timeline is the accumulation of respondent laboratories as each began HFE testing. The mean time from publication to adoption (of the truncated distribution) was 14 months. Significantly, 35 laboratories (60% of the 58 performing HFE testing at the time of our survey) reported introducing the clinical test before the first patent was issued.

We cannot say, on the basis of our data, whether the decrease in the rate of laboratory adoption after the date the patents were issued was because laboratory staff perceived that there was inadequate demand to justify their development of the test or

because of concerns about patent enforcement. Although our respondents overall reported that the patents weighed heavily in their decisions not to perform HFE testing, we do not know when those decisions were made, and that responses could have been biased by hindsight or by the nature of our questioning.

Our data show that there was very rapid adoption of HFE testing by laboratories soon after the cloning of HFE was published (ref. 8), roughly 17 months before the first patent was issued and almost 2 years before SBCL began enforcement. It is clear, as is typical with genetic tests, that the patents were unnecessary for rapid translation of the HFE discovery into clinical-testing services¹. On the contrary, our data show that the patents inhibited adoption, perhaps by creating a financial risk for laboratories, and a disincentive to develop and validate a clinical assay that could be stopped by patent enforcement.

Of course, without the potential value of the patented discovery, the investment of venture capital in Mercator Genetics might not have been made and the gene discovery delayed. Yet the question remains whether the exclusive licensing strategy embarked on by Progenitor and SBCL was the best method for capturing financial patent rewards in lieu of, for example, broad non-exclusive licensing of all laboratories that wished to perform the testing with payment of a reasonable per-test royalty, or development and sale of a test kit, the latter being the current strategy of Bio-Rad. But this last strategy may compromise test quality by restricting laboratories to using a single kit, thereby limiting innovation and development of alternative potentially higher-quality or lower-cost methods.

Counting costs

Our data highlight several concerns that have been expressed about patents on biotechnology discoveries. First, at least one study⁹ has shown that publication of new biotechnology discoveries is delayed because of patenting activities. The paper reporting the cloning of HFE (ref. 8) was submitted more than a year after the first US patent application was filed, and several months after the last of the four applications. Because laboratories can rapidly develop, validate and offer clinical tests, delay in publishing scientifically validated findings of clinical importance can adversely affect patients by delaying access to diagnostic testing.

Second, the two most common mutant alleles, covered by the US patents, account for up to 85% of haemochromatosis in the northern European population; there are many other rare polymorphisms that have clinical relevance¹⁰. Laboratories forced to use Bio-Rad's kit for financial reasons may decide to develop alternative tests for other mutations¹¹, which can increase the cost of testing, the likelihood of laboratory errors because of

Testing may be compromised by limiting labs to a single kit, as developing better or cheaper tests is not encouraged.

increased handling of samples, and, if any of the new mutations are also being patented, further increase the cost and licensing complexity for haemochromatosis testing.

Third, gene patents affect the cost and availability of clinical-diagnostic testing. Royalties charged for this and other non-exclusive licences include SBCL's and Bio-Rad's charge of up to \$20 per test (in addition to substantial up-front payments), and \$12.50 per test for Canavan's disease, \$5 per test for Gaucher's disease, and \$2 per test for volume greater than 750 tests a year for the most common allele ($\Delta F508$) of the CFTR gene that causes cystic fibrosis, all with no up-front fees. Although these amounts seem modest, they can present various problems. For example, 'stacking of royalties'¹² occurs for laboratories offering a panel of tests for the Ashkenazi Jewish population, including Tay-Sachs disease, several cystic fibrosis mutations, Gaucher's disease, Niemann-Pick disease and Canavan's disease.

One respondent indicated that his cost for this panel of tests was about \$100. The royalties for the tests make up about 20% of the cost; this percentage will increase as new tests are added to the panel and technology drives down the marginal cost per test. We believe that royalties must be reasonable and, given the rapid advances being made in testing technology, that they should not be fixed amounts but should be a percentage of the marginal reimbursement, cost or price that can be allocated to the patented test. ■

Jon F. Merz is at the Center for Bioethics, University of Pennsylvania, Philadelphia, Pennsylvania 19104-3308, USA; Antigone G. Kriss is at George Mason University School of Law, Arlington, Virginia 22201, USA; Debra G. B. Leonard is in the Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; and Mildred K. Cho is at the Center for Biomedical Ethics, Stanford University, Palo Alto, California 94304, USA. Requests for further data and information, including the full results of the survey, are available direct from J. F. M. (e-mail: merz@mail.med.upenn.edu)

1. Merz, J. F. *Clin. Chem.* 45, 324-330 (1999).
2. Wadman, M. *Nature* 413, 443 (2001).
3. Thomas, S. M., Davies, A. R. W., Birtwistle, N. J., Crowther, S. M. & Burke, J. F. *Nature* 380, 387-388 (1996).
4. Thomas, S. M., Birtwistle, N. J., Brady, M. & Burke, J. F. *Nature* 388, 709 (1997).



How much is clinical-diagnostic testing being limited by gene patents?

5. Gochee, P. A. & Powell, L. W. *Curr. Opin. Hematol.* 8, 98-104 (2001).
6. Schissel, A., Merz, J. F. & Cho, M. K. *Nature* 402, 118 (1999).
7. Cho, M. K. in *Preparing for the Millennium: Laboratory Medicine in the 21st Century* 47-53 (Am. Assoc. Clin. Chemistry Press, Orlando, Florida, 1998).
8. Feder, J. N. *et al. Nature Genet.* 13, 399-408 (1996).
9. Blumenthal, D., Campbell, E. G., Anderson, M. S., Causino, N. & Seashore, L. K. *J. Am. Med. Assoc.* 277, 1224-1228 (1997).
10. Pointon, J. J., Wallace, D., Merryweather-Clarke, A. T. & Robson, K. J. *Genet. Test.* 4, 151-161 (2000).
11. Thorstensen, K. *et al. Genet. Test.* 4, 371-376 (2000).
12. Heller, M. & Eisenberg, R. *Science* 280, 698-701 (1998).

Acknowledgements

Supported in part by the US NIH. The opinions expressed are those solely of the authors. We thank interview respondents; Elizabeth Miller for research assistance; and Michelle Henry and Meredith Weaver for comments. Preliminary results of this survey were presented in testimony to the US Congress in June 2000, and at the 10th Anniversary of ELSI Research Conference, Bethesda, Maryland, 17-19 Jan 2001.

The acceptable face of conservation

Wildlife conservation can bolster human needs rather than conflict with them.

**African Wildlife and Livelihoods:
The Promise and Performance of
Community Conservation**

edited by David Hulme

& Marshall Murphree

Heinemann: 2001. 344 pp. \$65

Brian Child

Community conservation in Africa ranges from park outreach programmes in East Africa to southern Africa's radical policies to give back to communities the rights to use and benefit from the wildlife on their land. The latter, in particular, have contributed significantly to the emerging idea that conservation should contribute to basic human needs rather than conflict with them. This radical revision of a paradigm built on 'fortress conservation' is already incorporated into key international conventions, yet its practice remains new and experimental. David Hulme and Marshall Murphree combine the experience of key practitioners with the insight of leading academics to produce the best analysis of community conservation available in print. The book is not a collection of disconnected case studies and analyses, but develops a well-crafted understanding of the concept.

Community conservation is being driven forward by a remarkably small number of 'scholar practitioners', who are attempting to realign institutions so that conservation benefits local communities and introduces democratic practices. Giving ownership of wildlife to local people, and encouraging profitable uses of it, provides a powerful incentive for conservation. 'De-nationalizing' wildlife by passing control to thousands of rural people raises a challenge. However, the real threat to the process is ineffective and predatory national leadership. Demographic growth and subsidized rural livelihoods add further pressure.

Community conservation is a fashionable topic, but the associated literature tends to be shallow and over-generalized, with few primary sources. Hulme and Murphree make the important observation that the real body of knowledge — insightful and aligned with reality through constant honing at the sharp end of conservation — is largely oral and inaccessible. In collecting together a sophisticated understanding of complex processes, the book demonstrates, in common with the business literature, that true progress depends on committed professionals actively tackling real problems in the field. Written academic knowledge often lags behind practice by at least a decade, and can become tangential to it. Edmund Barrow



Living side by side: a new idea of conservation focuses on maximizing the value of wildlife to African communities.

and Murphree's observation highlights the failure of conservation as a discipline. Its reward systems tend to attract talent away from the field, and it has a hegemonic structure that is strangely reminiscent of the Soviet command-and-control mentality, which does not lend itself to rapid innovation. To add to this, innovation is stultified because conservation agencies are not accountable for creating value, and because wildlife is treated as a national or global resource in a political market-place that is at the mercy of special-interest groups.

When conventional conservation failed to save the extensive wildlife outside protected areas, several African countries began to conduct experiments of market-led conservation with a strong human dimension. Even if these trials do not survive Africa's wider political turmoil, they are nurturing a new idea of conservation that focuses on maximizing the value of wildlife to landholders and communities. The most important results to come out of such work will be management principles and experience in carrying out the checks and balances that are needed for devolving the rights to wildlife. Although most community-conservation initiatives are not allowed to apply these principles, there is just enough evidence to indicate that when they are applied, the hoped-for synergy between conservation and development indeed occurs. Scale is critical, with the relative successes of village-level programmes confirming the argument that face-to-face interaction is an essential component.



The book develops East Africa's story using two case studies on community outreach in Uganda and Tanzania — although Kenya's experience, expensive and rather unsuccessful, is notably absent. But these initiatives lack the finances to be sustainable, and they should perhaps heed the lesson learnt by southern Africa that it is almost impossible to succeed if the economic signals are not taken into account.

The southern Africans have begun to revise the economic and political playing-field on which wildlife resources compete. They have made wildlife a profitable commodity that is owned by communities, and the book is particularly good at describing how this change came about. Three insightful case studies from Zimbabwe, Namibia and Mozambique illustrate the challenges of devolving management, and assess various aspects of the power relations. A deeper analysis is found in chapters on the incentives for institutional change (Bond), tenure and devolution (Murombedzi), and the evolution of participatory community wildlife management and monitoring (Taylor).

The result is a well-balanced, multi-

disciplinary presentation of the facts. Its only omissions are the importance of private conservation as a stepping-stone for devolved wildlife management, and why wildlife offers an economic solution in some of Africa's agriculturally marginal environments.

Southern Africa has collectively grasped the nettle, taking a leap that has revolutionary implications for conservation practice. Deeply committed scholar practitioners attempted to fast-track the empowerment of communities. But although it was initially highly successful, this strategy is now facing the challenge of 'aborted devolution', as politically motivated elite groups try to grab wildlife resources for themselves. The book offers the frank and honest observation that, although the underlying concepts appear to hold and to offer the best way forward, most community-conservation initiatives survive only through the efforts of rare individuals or by external financial support.

There is sufficient evidence that community conservation, when properly applied, works, and can do the greatest good for the greatest number. Yet proper application remains the exception rather than the norm. Its weakness lies in the assumption that the concept of the greatest good is a winning formula for bringing about institutional change. Particularly in non-democratic countries, there is much evidence to the contrary. ■

Brian Child is a conservationist based in Cape Town, South Africa.

A lament for Italy's brain drain

Cervelli in Fuga

sponsored by the ADI (Associazione Dottorandi e Dottori di Ricerca Italiani)
Avverbi: 2001. 189 pp. 9.29 euros

Roberto Battiston

There is a ghost wandering through Italy. It is a small, wasp-coloured book written by a group of brilliant young minds, born and educated in this beautiful country but living and working elsewhere. This disturbing book, sponsored by the ADI — the organization of Italian doctoral students and research doctors — is receiving a lot of attention and is to be found on the desks of an increasing number of academics, politicians and decision-makers throughout Europe.

The book, *Cervelli in Fuga* ('brains on the run'), describes the dark side of public research in Italy — its selection and promotion procedures, the corrupt foundations on which everything else is built. Over the past 30 years, periodical mass hirings or promotions in the research sector (the infamous "ope legis", that is, hiring or promoting by law) have lowered competitiveness. During the past five years, a new law regulating professorship promotions has practically ended the exchange of professors between Italian universities and research institutes. The effects of these distortions are far-reaching,

as shown by the recent scandal over the selection of directors for the newly formed National Research Council institutes — radical change in the appointment procedure had been promised, but the reality indicated 'business as usual'.

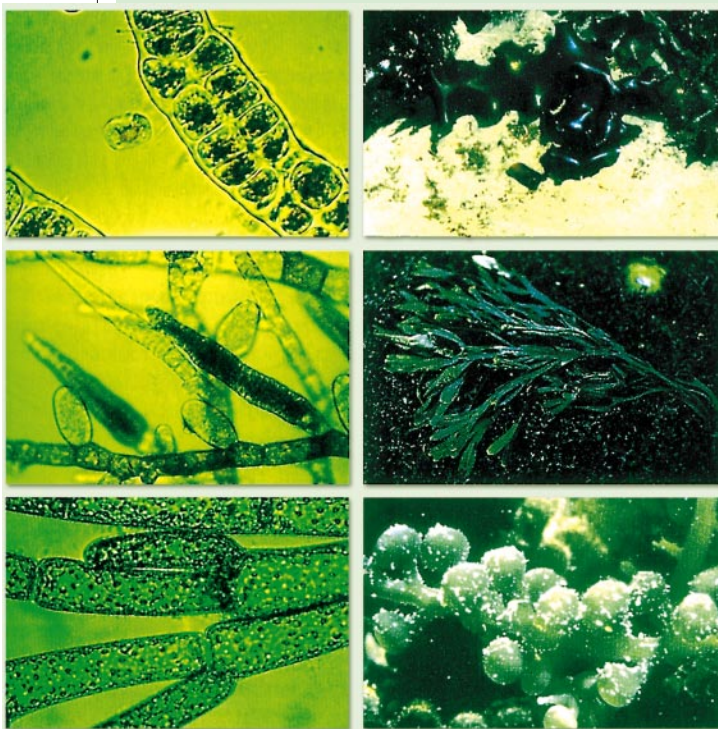
Every Italian knows about these problems, but this book gives a unique account of those directly affected which brings the message home powerfully. It explains, through 21 short autobiographical essays, why so many talented Italian scientists choose to work in other countries. The stories are honest and revealing.

Leaving a country is not intrinsically a bad thing to do in a world where everything is shared across continents, particularly science. But the message conveyed in these essays is one of humiliation. The authors, among the best minds of their generation, are strongly motivated and ready to take on the difficult life of academic research. But they say they had to leave to save themselves. They explain that it was not just a question of finding an academic position, but, more importantly, of retaining the feelings that are essential for a scientist. For example, Nicola Terrenato, a 37-year-old archaeologist who studied at the University of Rome and is now assistant professor at the University of North Carolina at Chapel Hill in the United States, deplores the fact that young scientists in Italy are under constant pressure to defer to their professors. He writes: "I am still moved when I find myself writing what I think, without indulging in concessions or unnecessary citations, and I realize that, had I remained in Italy, I would never have understood what I understand here."

Those who leave are often among the best. They are more motivated, but also less patient. Those who are good at merging into their boss's shadow, the 'portaborse', the yes-men, the academically cunning, those who learned from their family how to get on (Italian academic nepotism is a well-known plague), have a far higher chance of entering the system — a pernicious reverse-selection process described in a lively essay by Luca Ferrasin, a runaway research associate from a northern Italian university, now a contented lecturer at the Bristol Veterinary School in England.

The book also contains half a dozen contributions that analyse the position of the Italian university, its legal background and perspectives. In his essay "Inconclusive conclusions", Augusto Palombini, a doctoral student in archaeology and one of the main contributors to the book, summarizes Italy's "feudal culture", which has such ineffective selection procedures that the average age of an Italian research associate (the first level in the hiring process) is still today an appalling 45 years. Before blaming the lawmakers, he says, individual academics must bear some of the responsibility. But he admits that

Weeds' watery ways



The diverse size and shape of sessile marine organisms such as the seaweeds shown here are strongly influenced by their environment as well as their genome. *The Algorithmic Beauty of Seaweeds, Sponges, and Corals*, edited by Jaap A. Kaandorp and Janet E. Kübler (Springer, \$49.95, £35), describes how simulation models can throw light on the growth and structure of these organisms, with particular emphasis on seaweeds, sponges and corals.

things seem to be moving in the right direction, albeit slowly. For example, the level of funding for research has increased; in 2001, Giuliano Amato's government devoted an additional 450 million euros (US\$388 million) to research, using money from an auction of telecommunications licences — although this measure has not been repeated for 2002 by Silvio Berlusconi's government. Italy invests about one-third of the percentage of gross domestic product that Japan spends on research, although resources are starting to be distributed on a more rational basis, involving peer review. However, Palombini points out that simply putting more money into a system that contains too many people who do not share the same values of merit and competition is no guarantee of success.

Most Italian universities are still capable of producing highly talented scientists. Burton Richter, Nobel prizewinner and professor of physics at Stanford University in California, knows the Italian system well, being a member of an international committee charged with reviewing the Italian National Institute for Nuclear Physics. As he writes in this book, he does not understand how Italy can afford the social cost of this continuing brain drain, particularly in the absence of a brain gain from other countries, which is hampered by the bureaucratic and parochial selection system in Italy, as well as by poor funding. I don't understand it either. But reading this book gave me a better idea of the causes and consequences of this process. ■

Roberto Battiston is at the Sezione INFN, Via Pascoli, I-016121 Perugia, Italy.

The answer lies in the soil

Understanding Soil Change: Soil Sustainability over Millennia, Centuries, and Decades
by Daniel D. Richter Jr & Daniel Markewitz
Cambridge University Press: 2001. 272 pp.
£47.50, \$69.95

David Schimel

This marvellous little book tells the story of southeastern US ecosystems from the perspective of soil changes over timescales of decades, centuries and millennia. I was delighted with the author's organizing of processes into timescales. This is a natural way of thinking about soil and ecosystem processes, but is more often given lip-service than actually used as an organizing principle. In this case, it works, both because the soil processes — from weathering and geomorphic evolution, through leaching and decomposition — fall into this framework,

and because human impacts also change on these timescales.

The first part of the book provides a clear and conceptual overview of the elements of soil science, and their links to human management of soils. The authors demonstrate clearly that sustaining soil "health" is essential to sustaining human societies, and that soil change is linked to historical change in human economy. They then introduce the system that is the focus of their study, the Calhoun Experimental Forest, a property in South Carolina procured during the regional collapse of unsustainable agricultural practices of the southern United States in the first part of the twentieth century.

In the second section, which addresses millennial processes, they review pedogenesis (soil formation) globally and in the old soils of their target region. Here we begin to see the fruit of the sustained research effort at the Calhoun site, as the processes leading to the acid soils of the region are carefully documented in terms of geomorphic, geochemical and biological dynamics. Key findings include the role of biologically generated acidity in attacking the soil's primary minerals.

The third section addresses the centennial timescale. It includes a history of the agricultural development of the South, drawing attention to the little-known intensive maize cultivation system of the indigenous Americans before the arrival of Europeans. The authors provide a detailed analysis of the sustainability of this system, and the human and landscape conditions that allowed its success. They then provide a unique perspective on the 'Old South's' cotton economy in terms of agricultural management and the biogeochemistry of the cotton economy. The discussion on the legacy of cotton growth in today's soils makes fascinating reading, blending political and economic history with soil chemistry. The authors conclude that the previous two centuries of agriculture have affected soils to great depth (two metres) and will continue to affect ecosystems well into the future.

The authors next describe a period over recent past decades during which agricultural land over huge areas in the southeastern United States has reverted from farmland to forest. This is the timescale of the Calhoun experiment and describes in detail the recovery of carbon and nitrogen in soils of the regrowing forest. The rapid and massive response of soil to the re-establishment of forests is astonishing. The Calhoun soils re-establish soil acidity, carbon content and a functioning nitrogen cycle that can effectively retain inputs. Likewise, phosphorus chemistry changes as a result of the increase in organic matter and the associated effects of acidity on the inorganic chemistry of this element. The degree of knowledge of soil processes documented for



Down to earth: the Calhoun experiment has studied soil–ecosystem changes over 43 years.

this site is equalled for only a very small number of other locations.

The final section of the book outlines a research strategy based on a proposed network of long-term soil-research sites and careful experimental design. This proposal, in which the authors argue for a research paradigm based on replicated manipulative experiments, is founded on experience both at the Calhoun site and in other long-term ecosystem studies. While the utility, value and rigour of this approach need no comment, the authors themselves have shown the additional value of a perspective that takes into account the dynamics and legacies of millennial and centennial landscape processes. Decades-long experiments, using conventional manipulation and replicated plots, cannot address the role of slow pedogenic and landscape-scaled processes. For this, the analytical paradigms of the geophysical sciences could be useful, as they can address entities (such as ecosystems) for which replication is a challenging problem, which are difficult to manipulate, and which respond on long timescales. The dependence on fisherian statistics and classical experimental design in soil science is a legacy from agronomy, not altogether suited to large-scale, long-term processes.

This book fills an important niche in the biogeochemical literature, and not only as a regional case study. Forest ecosystems play a large part in global processes, affecting the

MICHAEL HOFMCKEL

global water, carbon and nutrient cycles. In global studies, processes must necessarily be captured in models, and in general, today's global models greatly simplify soil carbon dynamics, sketch out or ignore nitrogen and totally neglect phosphorus, acidity and cations. This study shows the importance of an integrated appraisal of soil dynamics in ecosystem function, and demonstrates the increasing maturity of soil science. ■

David Schimel is at the National Center for Atmospheric Research, Boulder, Colorado 80305, USA.

More from the soil

Tales from the Underground: A Natural History of Subterranean Life

by David W. Wolfe

Perseus, \$26, £18.99

Warfare of a chemical kind

War and Nature: Fighting Humans and Insects with Chemicals from World War I to Silent Spring

by Edmund Russell

Cambridge University Press: 2001. 336 pp.

£35, \$55 (hbk), £12.95, \$20 (pbk)

Alastair Hay

Metaphors are like pictures; they save text. So, when senior politicians refer to members of the al-Qaeda terrorist network as 'mosquitoes in a swamp', for example, we understand the allusion. An image is conjured up in our minds, and our response may be more visceral than cerebral. The use of metaphors about nature has a long pedigree and may have something to do with our historic contact with the land. Shakespeare, in *Richard II*, has Henry Bolingbroke (a future king) refer to those who wrongly occupy properties of his as: "The caterpillars of the Commonwealth, which I have sworn to weed and pluck away."

Military metaphors have a pedigree too, but their use in agriculture is much more a product of scientific developments in the twentieth century. As Edmund Russell points out in *War and Nature*, over the past 80 years, insects have come increasingly to be seen as the enemy, and suppliers of insecticides have used government and the press to encourage us to buy their products.

This approach was no more evident in the United States than in wartime, and in 1944 the Bureau of Entomology was urging dairy farmers to wage a "War on Insects". At the same time gardeners, if they saw insects, were ordered by the US Department of Agriculture to "shoot to kill". Even magazines for the home, such as *House and Garden*, were on side, insisting on an "all-out attack on fifth columnists in the garden".

By the Second World War, the use of such phrases was commonplace and did not jar with readers. The groundwork had been laid well and had started even before the end of the previous world war. As Russell points out in his book — which largely describes practice in the United States between 1917 and 1963 — towards the end of the First World War, thoughts were on future disarmament, and chemical munitions were high on the list of candidates. Those employed to make chemicals for war could see unemployment looming.

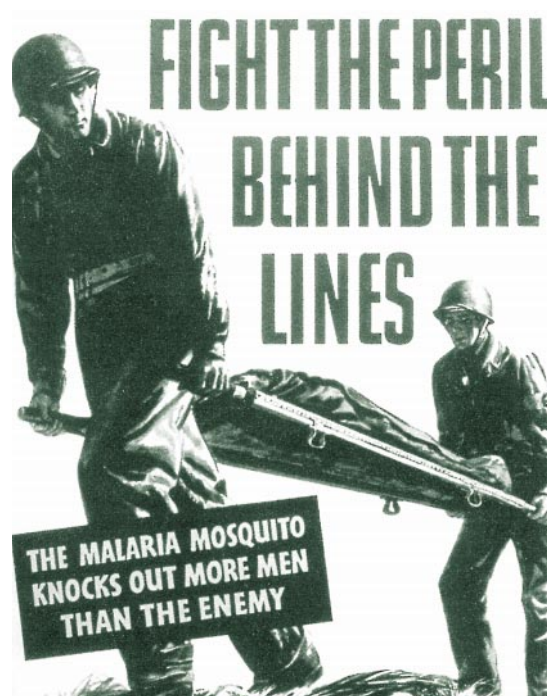
The fight by the US Chemical Warfare Service (CWS) to survive in a postwar era is well known by those who research this area. Arguments used by the agency and some of its leaders that chemical weapons were more humane than conventional munitions have become folklore. Analysis of the ratio of deaths to acute injuries certainly showed that chemical weapons caused far fewer fatalities than did bullets or explosives. Data on more chronic injuries were rarely referred to.

Proponents of chemical warfare also argued for even more toxic agents to be developed, so that their use could be threatened in future wars as a deterrent to an enemy. Today, it is not just the toxicity of modern chemical munitions, but their potential to cause significant civilian casualties, that is (hopefully) leading to their demise.

But it was not the development of more toxic chemical-warfare agents that helped the CWS to remain intact. The service's survival was due in large measure to some lateral thinking. Armed with the technology to disperse chemicals, all it needed was a new enemy. Step forward the boll-weevil. Long the bane of those who grew cotton, the boll-weevil could at that time only be controlled by the insecticide calcium arsenate. The CWS offered to find a better alternative.

Russell notes that in 1926, after a two-year research programme, alternatives were found, but none was more efficient than the arsenate. Although the outcome was not a success, press reports began increasingly to eulogize the research skills of the CWS. Government appropriations continued to flow to the agency, and collaboration with civilian government agencies and commercial firms followed. Survival of the CWS was assured by this approach and, apart from a later change of name to the Chemical Corps, the *modus operandi* of those involved with chemical warfare in the United States has remained the same.

Russell includes incendiaries as chemical-warfare agents. Germane though they are to his story about the CWS, the inclusion of incendiaries inflates the importance of chemical warfare in the Second World War. Chemical agents today are defined by their toxicity and their mode of action as chemicals that harm life, rather than their flammable properties. The CWS helped to devise



Airborne enemy: the malaria mosquito plagued troops in the Pacific in the Second World War.

and make many of the incendiaries used on Germany and Japan in the Second World War, causing great loss of life. Horrific beyond imagining, these deaths were due to the consequences of the fires and not to direct chemical poisoning.

Russell provides many examples of the interaction between industry and the military and its importance, such as the campaign to spray the insecticide DDT in the Pacific theatre during the Second World War to control the malaria mosquito then besieging US troops. Malaria caused five times as many casualties as military action.

Japanese troops and mosquitoes were juxtaposed in these campaigns, where the emphasis, both in pictures and text, was on elimination of the enemy. This, and other instances of the use of controlling chemicals, provides Russell with ample fodder for a thought-provoking and eminently readable book. A sequel might be about where much of it went wrong. ■

Alastair Hay is in the Molecular Epidemiology Unit, School of Medicine, University of Leeds, Leeds LS2 9JT, UK.

More on biological warfare

Britain and Biological Warfare: Expert Advice and Science Policy, 1930–65

by Brian Balmer

Palgrave, £45, \$75

Chemical and Biological Warfare: A Comprehensive Survey for the Concerned Citizen

by Eric Croddy, with Clarisa Perez-Armendariz & John Hart

Copernicus, \$27.50, £17

Coming to terms

Caloric, cathode, curium and quark — coinage from the mint of science.

J. L. Heilbron

Among the obstacles to the steady advance of science are the words invented to denote its conquests. Francis Bacon rated poor terminology as the most menacing of the four sets of idols he defined in his *Novum organum* of 1620 as befuddlers of the human mind. Badly coined words, the “Idols of the Market Place”, thoughtless phrases minted for the moment, “wonderfully obstruct the understanding”. The incautious wordsmith of science can “throw all into confusion, and lead men into numberless empty controversies”.

The history of scientific terminology opens a royal road to the history of scientific culture. The eighteenth century, which spent much of its intellectual energy classifying and summarizing its burgeoning knowledge, devised terminology that transformed botany and chemistry. The binomial designation of natural species introduced by Linnaeus and the systematic names of chemical compounds invented by Lavoisier and his collaborators remain in use, although they are not free from damaging idolatry. Linnaeus embedded his binomials in a system of arithmetically defined taxa that sometimes put species in the wrong families. The French chemists admitted the substance caloric, which does not exist, among their elements, and coined ‘oxygen’ on the mistaken idea that the gas so designated gave acids their acidity. But the terminology, erected on the enlightened principles of rationality, order and universality, proved flexible enough to drop erroneous reifications (like caloric) and ignore misnomers (like oxygen).

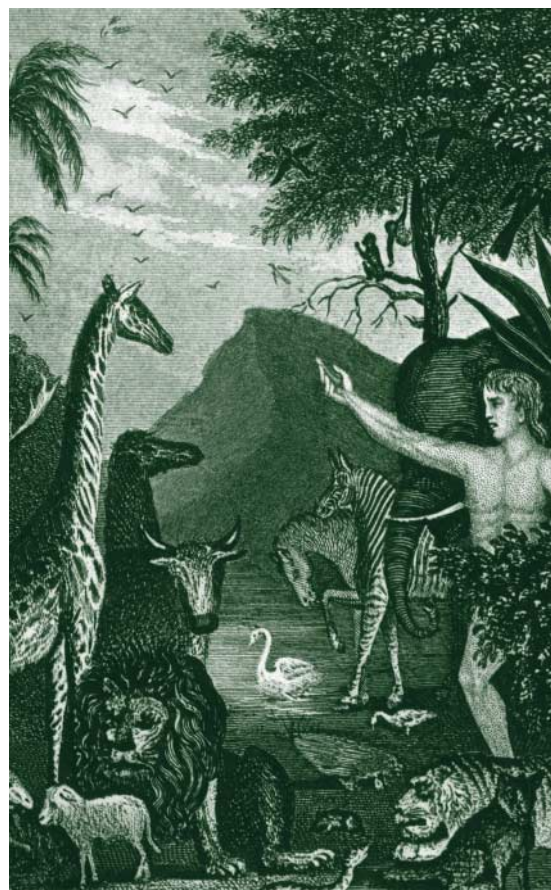
In the nineteenth century, geology, palaeontology and physics enlarged their vocabularies with the help of William Whewell, a veritable mint of new coinages. He did not try for grand systems, however, but for small groups of linked words that would convey concepts without fixing theories. In response to Michael Faraday’s

request for terms to describe his experiments in electrolysis, Whewell supplied anode, cathode, electrode and ion from non-committal Greek roots. He gave names to experimenters too, ‘scientist’ and ‘physicist’ for example; however, they had too much of the market-place about them (they came too close to ‘dentist’) for the ‘men of science’ they were supposed to designate.

Whewell’s criteria guided the choice of terminology in the established sciences until the First World War. His use of ancient languages and his conviction that the right words would establish correct and durable concepts fit his age perfectly. Most of the people whom he wanted to call scientists knew some Latin and Greek; and they and he built as if their constructions would be in use for ever.

In competition with Whewell’s rules of art, the later nineteenth century began the practice, in keeping with the growing nationalism of the time, of naming chemical elements and electrical standards after a discoverer, the country of the discoverer, or a national hero. In chemistry, the introduction of the new style can be traced to the 1870s and 1880s, when the discoverers of gallium, germanium and scandium named them after their nationalities (French, German, Swedish). During the twentieth century, nationalistic names attained the detail of a postal address (berkeleyum, californium, americium). The names of heroes, invoked to distinguish electrical standards during the late nineteenth century (ohm, amp, volt ...), invaded the periodic table a century later (curium, fermium, mendelevium ...).

After the Second World War, Americans gained by priority of discovery the right to name the elementary particles. Their terminology tended to be facetious and jocular. Thus, quarks in their various flavours and colours; gluons to paste quarks together; quantum chromodynamics, which does not study colour; and GUTs and TOEs, not body parts but Grand Unified Theories and Theories of Everything. Did the jocular indicate the easy confidence of people who felt close to finishing physics? It certainly demonstrated that the sober conservatism of European scientists of the nineteenth and early twentieth centuries had given way to the flippant equality of Americans during their time of world dominance. The playful names coined by high-energy physicists have been criticized as inelegant, non-ancient, capricious and misleading. No doubt it is unlucky that quark means garbage in German, but gluon



The first systematic naming was Adam’s task of finding names for the creatures God had made.

is an inspired put-on: it looks Greek, means nothing in German, puns in English and satisfies Bacon’s requirement that a word express a clear and distinct idea.

The earliest case of systematic naming occurred about 6,000 years ago, when God brought all his newly fashioned creatures to Adam, “to see what he would call them” (Gen. 2:19). That eases the task: one Provider of instances, one namer of names. Genetics and molecular biology have had a great many of both and, consequently, a taxing and awkward terminology. Students of fruitflies favour bouncy names in the style of particle physicists: *armadillo*, *hedgehog*, *lost-in-space*. Mouse geneticists like dull ones, such as *b-catenin*, which happens to be the same gene as *armadillo*. A single gene (*selectin L*) has 15 different aliases, whereas *MT1* refers to at least 11 different genes (see *Nature* **411**, 631–632; 2001).

The cure for this genetic disorder is a computer, which identifies a gene not by its name but by systematic descriptors. The new Adam condones local and multiple names for fundamental entities of science and undercuts the principle of univocal universality that has informed scientific terminology since the Enlightenment. A perfect post-modernist solution. ■

J. L. Heilbron is at Worcester College, Oxford OX1 2HB, UK.

The playful names coined by high-energy physicists have been criticized as inelegant, capricious and misleading.

A long engagement

Rod Flower

"To most modern pharmacologists," de Jongh wrote in 1964, "the receptor is like a beautiful but remote lady. He has written her many a letter and quite often she has answered these letters. From these answers the pharmacologist has built himself an image of this fair lady."

Although he was writing almost 20 years before the first detailed structural account of any receptor, at a time when our knowledge of receptor mechanisms was slight, this concept had already spawned many outstanding pharmaceutical innovations. One thinks, for example, of the development of beta-blockers and of selective histamine antagonists. So how did this most influential idea come about?

The notion that drugs act through receptors has its true origins in the late nineteenth century, when dogma held that the cell protoplasm is a single, very large, molecule. Paul Ehrlich introduced the term 'receptor' in 1900 to describe sites on this molecule to which bacterial toxins (and, later, drugs) bind to bring about changes in

cellular metabolism. The idea of discrete and specific binding sites was also supported by the experiments of John Newport Langley. His observations on the antagonistic actions of curare and nicotine on skeletal muscle eventually led to the key concept that drugs possess two properties: the ability to bind to the 'receptive substance' (as he called it) or affinity; and the ability of an agonist to cause an effect, later dubbed its efficacy. The final impetus came from Alfred Joseph Clark, who realized that the relationship between drug concentration and response could be described using a simple mathematical model, and who thus bequeathed us the rudiments of receptor theory.

By the late 1940s, the balance of evidence and opinion had swung towards the Ehrlich–Langley–Clark model of drug action, and in the ensuing years it was to become as close to a credo as science allows. These early pioneers had laid the foundations of experimental pharmacology and had ensured that even if its practitioners did not know what receptors actually were, they at least had the conceptual tools to work with them.

Originally, the term 'receptor' was applied generically to all drug targets because there was no clear sense of how binding gives rise to a biological effect. Some of these targets subsequently turned out to be enzymes or other molecules, and today the term 'receptor' is generally reserved for a molecule that acts as a biological signal transducer — usually for endogenous hormones or neurotransmitters.

Like the omnipresent mobile phone, a receptor must not only be able to pick out the correct signal from the blizzard of irrelevant traffic, but also cause it to be amplified and converted into a form that is useful to the recipient. Clearly, a mere docking molecule cannot accomplish such a feat. Langley, showing more than a little prescience, believed that the receptive substance "is capable of affecting the metabolism of the chief substance" (by which he meant the secretory, contractile or other apparatus) of the cell. A breakthrough in understanding how this could occur came with the discovery of signal-transduction systems — biochemical relays that link receptor occupation to the phosphorylation of key proteins, changes in calcium or accumulation of 'second-messenger' molecules that ultimately modify such cellular processes. Research into these phenomena has become a distinct discipline in its own right.

From the genome sequences of humans

Drug receptors

The term 'receptor' is generally reserved for a molecule that acts as a biological signal transducer, usually for hormones or neurotransmitters.

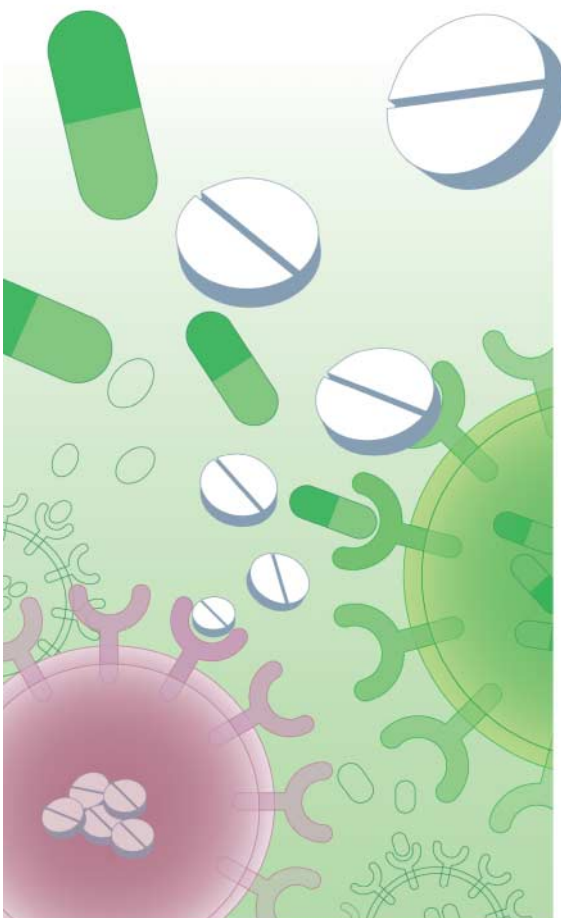
and other organisms, we now know that there are thousands of different cellular receptors. But we have also come to appreciate how thrifty nature has been in limiting their basic molecular design, providing just a handful of 'superfamilies'. The most abundant of these is the highly conserved seven-transmembrane-domain G-protein-coupled receptor, a design so adaptable that these receptors can be modified to detect 'ligands' as diverse as photons of light or the most complex polypeptide hormones. The opportunities for rational drug design on the basis of these similarities between members of superfamilies are not being overlooked.

But despite these tremendous advances, many questions remain. The nuances of drug action *in vivo*, structural data and predictions based on mathematical models can sometimes be irritatingly difficult to reconcile. Other disturbing facts have surfaced too — some receptors lack any signalling properties, whereas others seem to have their ligands already attached. Flouting conventional receptor theory, some drugs seem to behave both as antagonists and as agonists, depending on how they are administered; and (embarrassingly) even the fundamental idea of efficacy has been difficult to measure — or even to define. It seems that after 100 years of acquaintance and 50 of intellectual matrimony, our relationship with de Jongh's "mysterious lady" is still not fully consummated. Perhaps we should have insisted on a prenuptial agreement! ■

Rod Flower is at the William Harvey Research Institute, Charterhouse Square, London EC1M 6BQ, UK.

FURTHER READING

- Clark, A. J. *J. Physiol. (Lond.)* **61**, 530–546; 547–557 (1926).
 Langley, J. N. *J. Physiol. (Lond.)* **33**, 374–413 (1905).
 Parascandola, J. in *Discoveries in Pharmacology* Vol. 3 (eds Parnham, M. J. & Bruinvels, J.) 129–156 (Elsevier, Amsterdam, 1986).
 Clarke, W. P. & Bond, R. A. *Trends Pharmacol. Sci.* **19**, 270–276 (1998).
 de Jongh, D. K. in *Molecular Pharmacology* (ed. Ariens, E. J.) (Academic, New York, 1964).



VICKY ASKEW

Brains out of tune

Thomas F. Münte

Is lack of musical ability simply the result of failure to practise? Probably not, if new investigations are anything to go by. They show that a disorder akin to dyslexia affects the processing of pitch.

*There was an old fellow of Sheen
whose musical sense was not keen.
He said: "It is odd,
I can never tell 'God
save the weasel' from 'Pop goes the Queen!'"*

The old fellow of Sheen was in good company. Two presidents of the United States, Ulysses S. Grant (Fig. 1) and Theodore Roosevelt¹, were tone deaf, as was Che Guevara². Music educators often ascribe³ lack of musical aptitude to lack of practice. But in a paper published in *Brain*⁴, accompanied by a case report in *Neuron*⁵, Isabelle Peretz and colleagues suggest that 'congenital amusia' is a developmental disorder that can be placed alongside developmental dyslexia and specific language impairment.

The possibility that people can be subject to a selective, potentially inborn deficit of music processing has been entertained for more than a hundred years. But there have been only a few anecdotal reports to support such a notion. For example, the neuropsychologist Norman Geschwind described⁶ a man from a family with several musically impaired members who was fluent in three foreign languages yet could not sing a song, discriminate the pitch of two tones, or keep a rhythm, despite having had music lessons as a child.

Peretz and colleagues^{4,5} set out to put these observations on firmer grounds. To rule out a general learning disorder, they selected 11 subjects from a larger group of self-declared unmusical volunteers who had achieved a high level of education, had been exposed to music during childhood lessons, and had been unsuccessful in mastering music from the start. Previous research has shown that memory for songs and different aspects of music — such as interval (the pitch difference between two tones), contour (the general direction in which a melody is moving), rhythm (the local temporal structure of a piece) and metre (its global temporal structure) — can be distinguished

and are selectively impaired in patients with damage to particular parts of their brain⁷.

When their unmusical subjects were given a standardized battery of tests assessing these different dimensions, Peretz and co-workers found that there was a general failure in those tests involving pitch perception, whereas results in rhythm and metre tasks were more varied. In addition, the unmusical subjects were also insensitive to distortions of familiar tunes, and indifferent to dissonant chords. But they had no problem in recognizing familiar voices, environmental sounds and song lyrics. So a pitch-processing deficit was at the heart of the subjects' problems.

Peretz and colleagues next asked whether this deficit is specific to music, or whether it extends to other domains that depend on pitch information, such as language. In spoken language, crucial information is transmitted by modulating the pitch contour, or prosody. Consider, for example, the differences in intonation between "She is rich?" and "She is rich." Questions are indicated by a final rise in pitch of the order of six semitones, whereas for statements the pitch falls by about three semitones. The unmusical subjects had no difficulties in picking up these differences — that is, they could detect whether a stimulus was spoken as a question or as a statement — which suggests that the deficit is selective for music-like stimuli and spares the language domain.

What is the cause of this remarkable disturbance in pitch processing? Is it congenital, as Peretz and colleagues maintain? And if so, does it indicate the existence of a specific, genetically determined music-processing system? In support of the view that the deficit is hereditary, the authors point out that 6 of the 11 subjects described one of their parents, as well as some of their siblings, as unmusical, with other family members being said to be 'normal' musically.

These family histories hardly qualify as hard evidence. But fortunately it turns out that, as far back as 1925, geneticists found evidence for the heritability of pitch discrimination⁸. Last year, this issue was re-examined by Drayna *et al.*⁹ using state-of-the-art genetic modelling on a large sample of identical (monozygotic) and fraternal (dizygotic) twins in Britain. Such twin studies allow the influence of shared environments to be separated from that of shared genes. Drayna *et al.*⁹ presented subjects with a set of simple

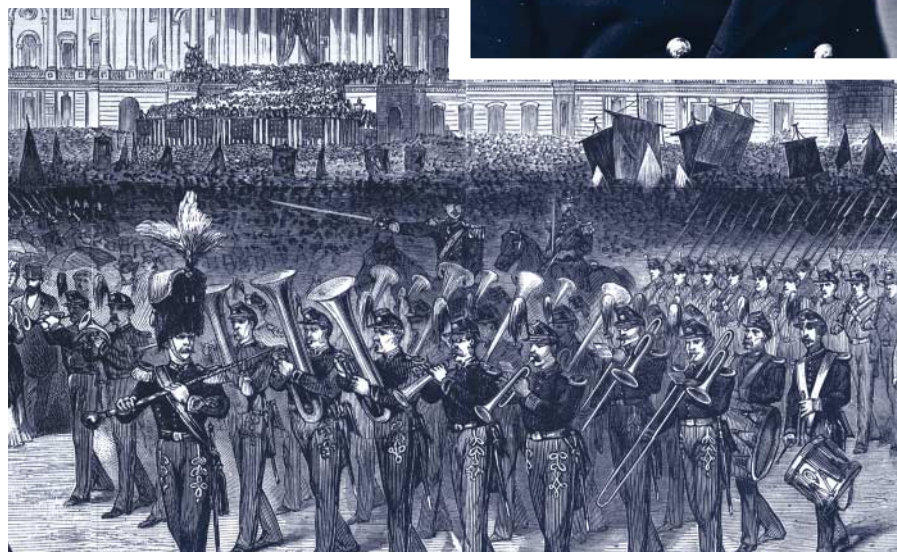


Figure 1 "I only know two tunes. One of them is Yankee Doodle and the other isn't." So said the tone-deaf Ulysses S. Grant (inset) to a reporter. Music annoyed Grant, so the band shown here, playing at his second inauguration in 1873, is unlikely to have set his toes tapping. Peretz and colleagues^{4,5} consider such 'congenital amusia' to be due to impaired processing of musical pitch.

CORBIS

BETTMANN/CORBIS



100 YEARS AGO

It is possible to believe that all the past is but the beginning of a beginning, and that all that is and has been is but the twilight of the dawn. It is possible to believe that all that the human mind has ever accomplished is but the dream before the awakening. We cannot see, there is no need for us to see, what this world will be like when the day has fully come. We are creatures of the twilight. But it is out of our race and lineage that minds will spring, that will reach back to us in our littleness to know us better than we know ourselves, and that will reach forward fearlessly to comprehend this future that defeats our eyes. All this world is heavy with the promise of greater things, and a day will come, one day in the unending succession of days, when beings, beings who are now latent in our thoughts and hidden in our loins, shall stand upon this earth as one stands upon a footstool, and shall laugh and reach out their hands amidst the stars.

H. G. Wells

From *Nature* 6 February 1902.

50 YEARS AGO

Studies have been made of the probable 'handedness' of prehistoric man by investigating the various relics of his tools and weapons. It appears from the way they are carved that prehistoric man was predominantly right-handed. The choice of one hand, usually the right but occasionally the left, and not either hand indiscriminately, is characteristic of man. It would seem that ambidexterity is an animal rather than a human characteristic... although to be left-handed in a right-handed society has numerous disadvantages, in spite of attempts to stamp it out in Britain, Greece, the United States and France, approximately 4–6 per cent of the population are still left-handed. Why should this be? Left-handedness appears to be inherited; how is not known. In certain families there seems to be a high incidence of left-handedness. Some investigators have found that such families also have more than the usual number of twins, so that there might be some connexion between twinning and left-handedness. That more males than females are left-handed seems to be agreed by investigators and raises the interesting question of whether this is the original distribution, or whether it is the result of the pressure of society.

From *Nature* 9 February 1952.

popular tunes, some of which had been altered by inserting a wrong note. The subjects differed greatly in their ability to detect the altered melodies, and genetic model-fitting indicated a heritability of pitch discrimination of 70–80%.

Does this mean, then, that the basic building-blocks of our music-processing system are to a large part inherited? Hundreds of pages have been filled over the past 150 years with attempts to explain the appearance of music, starting with Charles Darwin¹⁰, who believed that systems of calling in animals have a musical quality and evolved into speech. This view is echoed in a recent article by Gray *et al.*¹¹, which draws parallels between the songs of birds, whales and humans. Steven Pinker, the outspoken cognitive neuroscientist, on the other hand, notes that "of [all] mental faculties... music shows the clearest signs of not being [adaptive]"¹² and probably occurred as an epiphenomenon, as "auditory cheesecake".

The new data provided by Peretz *et al.*^{4,5} establish congenital amusia as a specific

developmental disorder. But to go further and help settle the debate, combined behavioural and genetic study of subjects with music-processing deficits will be needed to reveal more about the biological origins of our musical faculty.

Thomas F. Münte is in the Department of Neuropsychology, Otto-von-Guericke University, 39106 Magdeburg, Germany.

e-mail: thomas.muente@med.uni-magdeburg.de

- Williams, E. D. in *The American Experience* (TV programme produced for WGBH/Boston, 1996). Transcript available at www.pbs.org/wgbh/amex/presidents/nf/resource/tr/trscript.html
- Taibo, P. I. II *Guevara, Also Known as Che* (St Martin's Press, New York, 1999).
- Welch, G. F. *Res. Studies Music Educ.* **11**, 27–41 (1998).
- Ayotte, J., Peretz, I. & Hyde, K. *Brain* (in the press).
- Peretz, I. *et al. Neuron* **33**, 185–191 (2002).
- Geschwind, N. *Ann. Dyslexia* **34**, 319–327 (1984).
- Schuppert, M., Münte, T. F., Wieringa, B. M. & Altenmüller, E. *Brain* **123**, 546–559 (2000).
- Mjøen, F. *Hereditas* **7**, 161–188 (1925).
- Drayna, D., Manichaikul, A., de Lange, M., Snieder, H. & Spector, T. *Science* **291**, 1969–1972 (2001).
- Darwin, C. *The Descent of Man, and Selection in Relation to Sex* (Appleton, New York, 1871).
- Gray, P. M. *et al. Science* **291**, 52–54 (2001).
- Pinker, S. *How the Mind Works* (Norton, New York, 1997).

Microbiological oceanography

Hidden in a sea of microbes

David M. Karl

The photosynthetic activities of bacteria in the oceans are more diverse than previously thought. A full picture of the marine energy budget will require their separate contributions to be teased apart.

Green plants have been using oxygenic photosynthesis, in which oxygen is released, for more than 3 billion years. But there are two other non-oxygenic photosynthetic pathways, used not by green plants, but by certain bacteria. One pathway is known as anaerobic anoxygenic photosynthesis (AnAnP), because it can occur in the absence of oxygen; it pre-dates oxygenic photosynthesis, and is nowadays restricted to a few groups of bacteria that inhabit sunlit, oxygen-free habitats. The other pathway — aerobic anoxygenic photosynthesis (AAnP) — requires oxygen but does not generate it as a by-product; this pathway was discovered in marine bacteria only twenty years ago¹.

Most photosynthetic microorganisms in the open ocean were thought to be oxygenic, but there is growing evidence^{2,3} that oxygen-consuming, light-harvesting AAnP bacteria could make up as much as 11% of the total marine community. On page 630 of this issue, Béjà *et al.*⁴ identify new groups of AAnP bacteria in the sea, and show that these are much more diverse than expected.

All of these various photosynthetic bacteria harvest light energy using specialized pigments (the 'photo' part) and can convert CO₂ into organic carbon (the 'synthesis'

part), albeit with differing efficiencies. This is why the oceans are thought to act as carbon sinks: marine bacteria and algae convert atmospheric CO₂ into organic matter, some of which then enters the food web, where it remains for variable time periods. In oxygenic photosynthesis, chlorophyll *a* is the primary pigment responsible for harvesting light energy, and water is the hydrogen donor for CO₂ reduction, so oxygen is generated as a by-product (Fig. 1). Much of the oxygen will be consumed by non-photosynthetic organisms during respiration, when they metabolize organic matter to generate energy and to synthesize cellular constituents.

Before the dawn of oxygenic photosynthesis, AAnP bacteria used hydrogen sulphide or hydrogen gas, which were both abundant, as the hydrogen donors — this is why they do not produce O₂ (Fig. 1). These bacteria use bacteriochlorophyll *a* (Bchl_a) as the pigment for harvesting light energy. AAnP bacteria, on the other hand, use oxygen to metabolize organic carbon, to synthesize Bchl_a for example, but do so more efficiently when sunlight is available (Fig. 1). Like AAnP bacteria, they do not use water as the hydrogen donor, so oxygen is not produced. These bacteria are being found in surprising amounts in the open sea^{2,3}, and could make

us rethink our picture of carbon and energy flow in the oceans.

Béjà *et al.*⁴ analyse genes encoding the photosystem and photosynthetic pigments from bacteria collected in the northern Pacific Ocean (both open sea and coastal sites), and report an “unsuspected diversity” among marine AAnP bacteria. A novel aspect of this study is the use of molecular techniques — specifically the polymerase chain reaction after reverse transcription of RNA — to identify active ‘photosynthetic genes’ in AAnP bacteria that have never been cultured in the laboratory, including several new α -proteobacterial species not previously found in the sea. These culture-independent bacteria discovered by genomic analysis are genetically, and perhaps physiologically, distinct from known AAnP cultures. In the past, the discrepancy between the number of bacteria that are easily cultured and the abundance of microorganisms found in ocean samples has been a source of frustration and uncertainty, so genomic studies of this sort are likely to become increasingly important in the future.

The discovery⁴ of wide diversity among AAnP bacteria in the sea is important for several reasons. First, our understanding of the biogeochemistry (such as the carbon cycle) and the ecology (including the food web) of marine ecosystems is based on an oxygenic photosynthesis model typical of green plants. But local rates of photosynthesis in selected marine ecosystems may not always be linked to oxygen dynamics, and this may help to resolve a controversy about whether the open ocean is a net producer or a consumer of oxygen. If AAnP bacteria are important contributors to total marine production of organic matter, this would tip the apparent oxygen balance towards net consumption in their favoured habitats.

Second, Béjà and colleagues’ work⁴ follows closely on the heels of other reports of new marine microorganisms, including *Prochlorococcus marinus* (the dominant oxygenic photosynthetic organism in the open ocean⁵), planktonic *Archaea* in both surface⁶ and subsurface⁷ marine waters with yet unknown metabolic or ecological function, proteorhodopsin-containing bacteria⁸ that may be able to metabolize light and organic matter simultaneously, and new unicellular nitrogen-fixing cyanobacteria⁹ and small eukaryotes¹⁰. Collectively, these previously unknown marine microbes and unexpected metabolic diversity — now revealed by developments in molecular-based methods — are demanding a revision of our basic concepts of carbon and energy flow in the sea (Fig. 1). The potential ecological impact of these discoveries may be on a par with that of the original discovery of “little animals” (bacteria) in the sea by van Leeuwenhoek¹¹ more than three centuries ago.

We must now determine the physiologi-

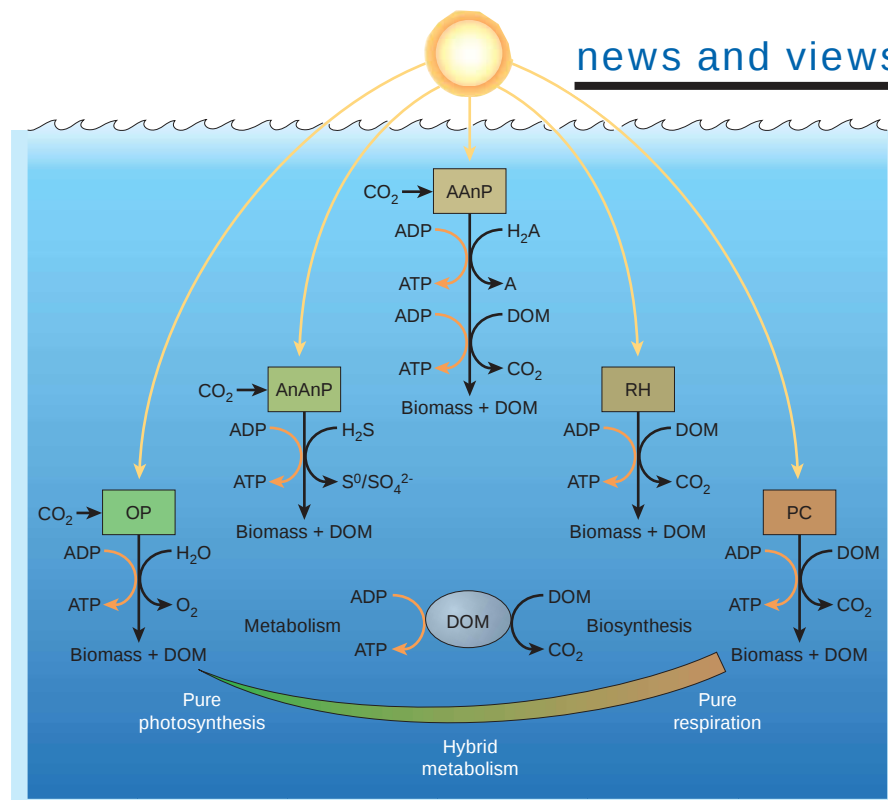


Figure 1 A current view of the complex relationships between sunlight, biological energy production and dissolved organic matter (DOM) in the open sea. Recent discoveries of new marine bacteria^{2–9} include all three modes of photosynthesis: oxygenic photosynthesis (OP), anaerobic anoxygenic photosynthesis (AAnP) and aerobic anoxygenic photosynthesis (AAnP), as well as two other potentially important light-driven processes, rhodopsin-based (RH) and phytochrome-based (PC) interactions that involve both light and DOM. Together, these light-driven processes, as well as others not shown here, sustain and control the flow of external energy into the global ocean. Each metabolic pathway is also dependent on the availability of DOM, ranging from low dependence (true photosynthesis) to high dependence (light-stimulated DOM respiration). DOM has at least two key functions: metabolism (ATP formation) and biosynthesis. Pure OP may be the exception in low-nutrient oceans, whereas mixed light–DOM metabolic processes are more likely in the open sea. Ecologists have yet to establish a comprehensive metabolic budget for these complex marine systems.

cal, metabolic and ecological relevance of each new group of bacteria to the ocean ecosystem. Goericke¹² recently reported that the percentage of AAnP pigments found in bacteria increased as total pigment content decreased, so the pigment-poor open sea may be a dominant habitat for AAnP bacteria. That study downplayed the role of AAnP bacteria in total photosynthesis because of the low overall occurrence of Bchl_a, but AAnP bacteria are known to contain much less pigment per unit biomass than oxygenic bacteria and algae because they have other metabolic options and so rely less on light energy¹³. It is possible, even likely, that the newly discovered AAnP bacteria can use light and organic matter simultaneously. However, without complete knowledge of their metabolism, a comprehensive energy budget for the marine ecosystem will remain missing from the global ecological puzzle.

Future studies must recognize and embrace the fact that a comprehensive understanding of the marine ecosystem is literally hidden in a sea of microbes. The diversity of microorganisms in the marine environment and the broad spectrum of their metabolic potential, including gene expression and regulation, are rapidly becoming as large

as the sea itself. By most accounts, we can culture fewer than 10% (by number) of the microbial inhabitants of the sea, so existing ideas of marine ecology must be flexible and accommodating to change. Both the stakes and the level of excitement in microbiological oceanography are at an all-time high, and the field, in my view, is poised for even more significant discoveries in the near future.

David M. Karl is in the School of Ocean and Earth Science and Technology, University of Hawaii, Honolulu, Hawaii 96822, USA.

e-mail: dkarl@soest.hawaii.edu

- Shiba, T., Simidu, U. & Taga, N. *Appl. Environ. Microbiol.* **38**, 43–45 (1979).
- Kolber, Z. S., Van Dover, C. L., Niederman, R. A. & Falkowski, P. G. *Nature* **407**, 177–179 (2000).
- Kolber, Z. S. *et al. Science* **292**, 2492–2495 (2001).
- Béjà, O. *et al. Nature* **415**, 630–633 (2002).
- Chisholm, S. W. *et al. Nature* **334**, 340–343 (1988).
- DeLong, E. F. *Proc. Natl Acad. Sci. USA* **89**, 5685–5689 (1992).
- Karner, M. B., DeLong, E. F. & Karl, D. M. *Nature* **409**, 507–510 (2001).
- Béjà, O. *et al. Science* **289**, 1902–1906 (2000).
- Zehr, J. P. *et al. Nature* **412**, 635–638 (2001).
- Moon-van der Staay, S. Y., De Wachter R. & Vaulot, D. *Nature* **409**, 607–610 (2001).
- Van Leeuwenhoek, A. *Phil. Trans. R. Soc. Lond.* **11**, 821–831 (1677).
- Goericke, R. *Limnol. Oceanogr.* **47**, 290–295 (2002).
- Yurkov, V. V. & Beatty, J. T. *Microbiol. Mol. Biol. Rev.* **62**, 695–724 (1998).

DNA repair

An APE that proofreads

Josef Jiricny

An enzyme involved in one aspect of DNA repair is now shown to have another function that improves the accuracy of repair. The new findings should help to improve therapies for HIV.

Base excision repair¹ is one of the ways in which damaged DNA is fixed. The enzyme DNA polymerase β has an important role in this process, but it tends to make mistakes. In their paper on page 655 of this issue¹, Chou and Cheng show how those mistakes are corrected. This in itself is striking. But their research also provides pointers to improving the effectiveness of HIV treatments by making them less toxic to the patient.

The job of base excision repair is to remove damaged bases from genomic DNA, as well as the sugar–phosphates that remain after the loss of bases². The process is outlined in Fig. 1. The trouble is that polymerase β , which is responsible for inserting the correct nucleotide, is embarrassingly error prone, making on average one mistake per 4,000 base insertions³. Assuming that between 20,000 and 80,000 base modifications take place daily in our genome⁴, polymerase β would be expected to introduce around ten mutations into our DNA each day.

Why this does not happen has long been a puzzle, but Chou and Cheng¹ have come up with a plausible answer. They show that the human apurinic or apyrimidinic (AP) nuclease APE1, an enzyme long known to function upstream from polymerase β in base excision repair (Fig. 1), and to interact with it physically⁵, can also act as its proofreading exonuclease by removing wrongly inserted nucleotides from the repair patches. In other words, an enzyme that normally acts by cleaving the DNA in intact strands, by hydrolysing the phosphate groups joining two sugar residues, can also act as an exonuclease that ‘nibbles’ off nucleotides from DNA ends that do not terminate in correct base pairs.

DNA bases can be damaged by many agents⁴. Some damaged bases are lost from DNA spontaneously; others are removed enzymatically by DNA glycosylases (Fig. 1). In both cases the result is an AP site in DNA, the removal of which is initiated by AP endonucleases, known as APEs. These enzymes cleave the sugar–phosphate backbone at AP sites to generate a free 3′OH terminus upstream from the site of base loss. The sugar–phosphate fragment that remains transiently attached to the 5′ terminus of the incised strand is then removed by polymerase β when the missing nucleotide is inserted⁶.

In cases where the polymerase inserts the right nucleotide — that is, one that forms

the correct (Watson–Crick) base pair with the nucleotide in the other strand — another enzyme, DNA ligase, completes the process by sealing the remaining nick². But polymerase β often inserts the wrong nucleotide³. In this case, the ligase cannot seal the remaining nick because the 3′OH and the 5′phosphate termini are misaligned.

In the bacterium *Escherichia coli*, the polymerase (pol I) that acts in base excision repair is different, and misinsertions do not pose a problem. That’s because pol I has a 3′→5′ proofreading exonuclease activity that can remove the mispair and resynthesize the correct nucleotide. But polymerase β has no associated proofreading activity, so it cannot catalyse the removal and replacement of the wrong nucleotide.

Chou and Cheng¹ now show that a nicked double-stranded DNA, with a mispair on the 3′ side of the strand discontinuity, is the best substrate for the 3′→5′ exonuclease activity of APE1. This activity decreased when a substrate with a single nucleotide gap was used, and diminished further with a bigger gap. It seems likely, then, that APE1 has a proofreading function in base excision repair.

The APE1 enzyme has been known for

many years and has been studied in detail by several groups. So how did its proofreading ability escape detection until now? The most likely answer is simply that no one looked for it, as there was no reason to suspect that it had this second function. The 3′→5′ exonuclease action of APE1 had been identified, along with other activities⁷. But the exonuclease action was much weaker than the endonuclease activity, and it was thought unlikely to be biologically significant. Moreover, the enzyme was known to act upstream from polymerase β , where there is no need for an exonuclease, so there seemed to be no compelling reason to test it on a substrate that might arise through erroneous gap-filling by polymerase β .

Chou and Cheng’s discovery¹ was fortuitous, because they had originally tackled a different task. They set out to identify a detoxifying enzyme that removed deoxyribonucleoside drugs from the DNA of leukaemic cells. (Leukaemias should be particularly sensitive to this type of therapy, because the cells affected proliferate rapidly and incorporate much more of these toxic drugs into their DNA than do other human cells.) When the detoxifying enzyme turned out to be APE1, the link with base excision repair was established, and it was only a short step to ask what the physiological substrates of this potent 3′→5′ activity might be.

But the authors did not forget the link of APE1 to the processing of nucleotide analogues. They also show that this proofreading exonuclease efficiently removes AZT and D4T (chain terminators used to inhibit HIV replication) from the 3′ termini

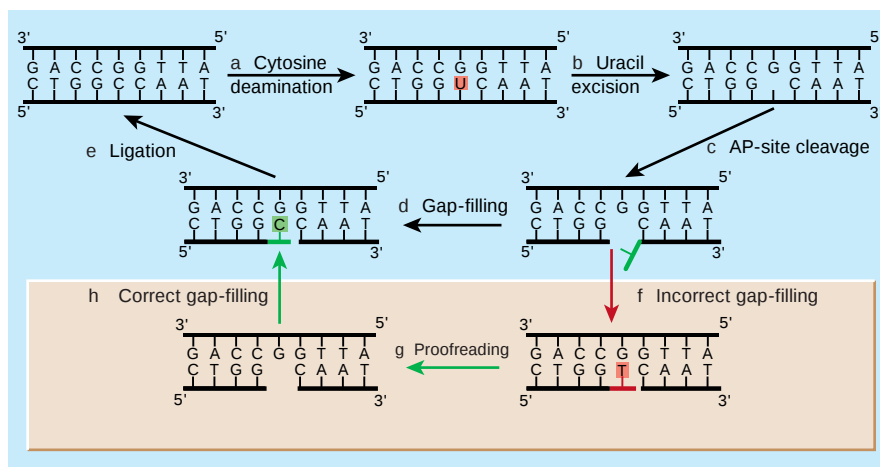


Figure 1 Base excision repair in human cells. **a**, In this example, a uracil residue has arisen spontaneously from cytosine through deamination, and needs to be replaced by another cytosine. **b**, The uracil is excised by the enzyme DNA glycosylase to generate an apurinic or apyrimidinic (AP) site, which in turn (**c**) is cleaved at its 5′ end by the AP endonuclease APE1. The repair process can then take one of two routes. **d**, The resulting single-nucleotide gap is filled in by polymerase β , which incorporates the dCMP-nucleotide carrier of cytosine, and also removes the sugar–phosphate flap (shown hanging from the lower strand). **e**, To complete successful repair, the remaining nick is sealed by a DNA ligase enzyme (ligation). The alternative route **f–h** occurs when polymerase β errs, and the events identified by Chou and Cheng¹ occur. **f**, dTMP, the carrier of thymine, has been erroneously inserted into the gap. **g**, It is recognized and removed by the 3′→5′ proofreader, APE1. **h**, Polymerase β then correctly fills the gap by incorporating a dCMP residue, after which ligation completes the process as normal.

of nicked DNA substrates. This finding could be of therapeutic importance.

Dideoxynucleoside drugs and their analogues — such as ddC, ddI, AZT, D4T and T3C — are not usually incorporated into genomic DNA in large amounts. But they do get into our genome through the action of polymerase β , which can incorporate them into repair patches during base excision repair⁸, and through another polymerase (γ), which inserts them into mitochondrial DNA⁹. Both of these events should cause problems, as incorporating these substances into DNA generates strand breaks that cannot be joined. Thus, by removing them from genomic DNA, APE1 helps our cells to survive.

In contrast, because DNA synthesis by HIV takes place in the cytoplasm, where there is no APE1, chain terminators incorporated into the proviral DNA by the HIV enzyme reverse transcriptase are not removed. As a result, synthesis of full-length HIV DNA can be inhibited. So, by improving the efficiency of APE1-mediated removal of chain terminators from genomic and mitochondrial DNA, it should be possible to reduce the side effects of HIV therapy.

The APE1-catalysed removal of nucleoside analogues from DNA termini is also interesting from the structural point of view. All of these substances form standard Watson–Crick pairs with the nucleosides in

the complementary strand. This implies that, rather than seeing the mispairs as such, APE1 may be detecting subtle distortions of the double helix at the nick, characterized by misalignments of the strand ends that make it impossible to rejoin them.

Gorillas that have learned to use sign language remain a rarity, even though they pop up on television from time to time. The newly discovered proofreading APEs, by contrast, are clearly commonplace and, in contrast to signing gorillas, likely to be of considerable clinical importance. ■

Josef Jiricny is at the Institute of Medical Radiobiology, University of Zürich and the Paul Scherrer Institute, August Forel-Strasse 7, CH-8008 Zürich, Switzerland.

e-mail: jiricny@imr.unizh.ch

1. Chou, K.-M. & Cheng, Y.-C. *Nature* **415**, 655–659 (2002).
2. Lindahl, T. *Mutat. Res.* **462**, 129–135 (2000).
3. Osheroff, W. P. et al. *J. Biol. Chem.* **274**, 3642–3650 (1999).
4. Lindahl, T. *Nature* **362**, 709–715 (1993).
5. Bennett, R. A. et al. *Proc. Natl Acad. Sci. USA* **94**, 7166–7169 (1997).
6. Sobol, R. W. et al. *Nature* **405**, 807–810 (2000).
7. Wilson, D. M. & Barsky, D. *Mutat. Res.* **485**, 283–307 (2001).
8. Wiebauer, K. & Jiricny, J. *Proc. Natl Acad. Sci. USA* **87**, 5842–5845 (1990).
9. Feng, J. Y. et al. *J. Biol. Chem.* **276**, 23832–23837 (2001).

Granular materials

Taking the temperature

Bob Behringer

The ‘temperature’ of a granular material depends on its entropy, but is hard to measure in the laboratory. So a theory that ties temperature to grain mobility and diffusion is welcome.

Granular materials are collections of solid, macroscopic particles, characterized by a loss of energy whenever the particles interact¹ (for instance, through friction when grains collide). On page 614 of this issue², Makse and Kurchan offer new insight into the statistical properties

of granular materials. In particular, their simulations suggest ways of determining a granular ‘temperature’, defined by the kinetic motion of the grains in much the same way that the random motion of molecules in a gas defines its temperature³.

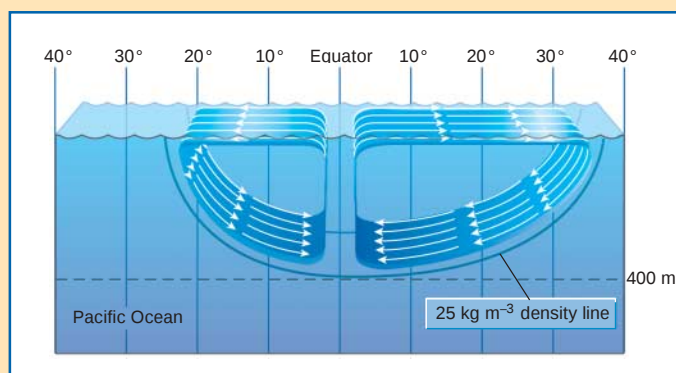
Typical granular materials include coal,

Oceanography

A slower flow

The deep ‘overturning’ circulation in the North Atlantic, in which northward-flowing surface water sinks at high latitudes and flows back at depth, may not be fully understood, but its existence has long been public knowledge. The Pacific has held on to some of its secrets for longer. Only in the 1990s were two shallow overturning circulation cells discovered, lying on either side of the Equator. As they describe elsewhere in this issue (*Nature* **415**, 603–608; 2002), M. J. McPhaden and D. Zhang have now produced an analysis of historical data to show that both cells have been slowing since the 1970s.

The circulation loops occur because winds produce surface currents away from the Equator. Water masses then flow along surfaces of constant water density (the figure shows the 25 kg m^{-3} surface as an example) from subtropical regions towards the Equator, at depths of 100–400 m. Near the Equator, upwelling water



replaces that lost at the surface to wind action.

McPhaden and Zhang have looked at data for the Pacific between 20° S and 50° N for the years 1950 to 1999. The calculated volumes of water transported in both circulation loops were much the same for 1956–65 and 1970–77, but declined rapidly in ensuing years. In the period 1990–99, the circulation volume was 25% less than at the earlier times.

This slowdown could help

account for other changes that occurred in the past couple of decades. For instance, sea surface temperatures in the tropical Pacific Ocean have risen by about 0.8 °C since the 1970s, despite increasing cloudiness. And the amount of carbon dioxide released into the atmosphere from that part of the ocean has declined. Both observations can be explained by a reduction in the delivery of relatively cool, CO₂-rich subtropical water to the equatorial ocean.

Perhaps the most intriguing question is whether there is any connection here with the El Niño/Southern Oscillation (ENSO), which occurs on shorter timescales. A shift towards stronger and more frequent El Niño events occurred in 1976–77, just when the decadal decline in the circulation began. But it is not clear whether the longer-term variation in circulation is a consequence of the increasing number of El Niño years, or whether the circulation slowdown provided the background for the change in ENSO cycles.

In the North Atlantic, changes in atmospheric conditions (the so-called North Atlantic Oscillation index) have brought milder and stormier winters to Northern Europe. As with the events documented by McPhaden and Zhang, this is a change that occurs over a period of several decades. There may in both cases be a connection with human-induced climate change, but for the moment that possibility remains speculative.

Heike Langenberg



BILL BACHMAN/SPL

Figure 1 New work² clarifies the thermodynamics of granular materials, such as sand.

wheat, sand (Fig. 1) and all sorts of powders. The study of such materials has a long history dating back to at least the time of Coulomb⁴, whose friction law was actually derived for these materials rather than for blocks moving on planes. It was Reynolds⁵ who first showed that compacted materials must first expand (dilate) if they are to deform. And Faraday's observation⁶ of convection in a shaken powder provided an insight into the role played by the surrounding air. More recently, investigations of granular materials have been driven by commercial applications. Vast sums of money are expended in handling these materials for applications that range from energy production or extraction to food supplies and pharmaceuticals.

In the past few years, researchers have taken up the issue of the statistical properties of granular materials. Because these materials have flow characteristics that roughly resemble those of ordinary, newtonian fluids, it is tempting to look in that direction for useful analogies. But these analogies should not be pursued too closely: granular systems dissipate energy quickly, so ordinary techniques of statistical mechanics that depend on energy conservation break down.

Undeterred, several groups have worked towards understanding the true statistical properties of granular materials. Such work includes studies of gas-like granular states⁷, investigations of fluctuations and stress variability in dense systems⁸, and, perhaps most importantly here, proposed new versions of statistical mechanics that would apply in a granular system where energy is not conserved^{9–12}.

This issue of energy conservation was addressed early on by Edwards and co-workers^{9,10}. They introduced a new way of calculating the entropy (the degree of disorder) and temperature for a granular system. Entropy depends on the number of states available to a system. Edwards *et al.* counted the number of configurations possible for fixed volume and energy, and defined the 'Edwards temperature' in terms of the volume derivative of this entropy.

The problem with the Edwards approach is that it is difficult to test directly in real

systems. Some experiments have achieved at least a flavour of the Edwards picture, such as studies of granular compaction^{12,13} that showed very slow increases in the density of a granular column that is tapped repeatedly. But, as with many granular experiments, their interpretation is complicated because there is a non-uniform rate of energy injection and non-uniform density — the key control parameter.

In Makse and Kurchan's numerical model² of a granular system subject to a gentle shearing force, the rate of energy input is uniform (over a reasonable time scale), and a uniform density is maintained. The authors use a tried and tested means to extract a temperature for the system: a comparison of diffusivity and mobility, whose ratio yields the temperature in conventional fluids. They then show that, for their model granular system, this ratio is identical for different particle sizes, just as one would expect in a fluid. The temperature extracted from their observations of diffusion and mobility is also consistent with the predicted value of the Edwards temperature.

So this work shows that, at least in a model system, it is possible to relate temperatures obtained from transport measurements, such as diffusion and mobility, to the Edwards picture. If similar transport measurements are performed in real granular systems, we would have, arguably for the first time, a measurement of a key statistical property that could be matched, in principle, to a theoretical prescription.

But there is a caveat: when real granular systems are sheared or shaken, they tend to become inhomogeneous often in both space and time. Although, to a certain extent, inhomogeneities can be avoided or accounted for, the link to the theoretical picture would be broken. Nonetheless, this work² leads us towards a clearer understanding of the statistical properties of real dense granular materials. ■

Bob Behringer is in the Department of Physics, Center for Nonlinear and Complex Systems, Duke University, Durham, North Carolina 27708-0305, USA.

e-mail: bob@phy.duke.edu

- Jaeger, H. M., Nagel, S. R. & Behringer, R. P. *Rev. Mod. Phys.* **68**, 1259–1273 (1996).
- Makse, H. A. & Kurchan, J. *Nature* **415**, 614–617 (2002).
- Jenkins, J. T. & Savage, S. B. *J. Fluid Mech.* **130**, 187–202 (1983).
- Coulomb, C. *Memoir de Mathématique et de Physique*, Vol. 7 p. 343 (Académie des Sciences, l'Imprimerie Royale, Paris, 1773).
- Reynolds, O. *Phil. Mag.* **20**, 469 (1885).
- Faraday, M. *Phil. Trans. R. Soc. Lond.* **52**, 299 (1831).
- Rouyer, S. R. & Menon, N. *Phys. Rev. Lett.* **85**, 3676–3679 (2000).
- Howell, D., Behringer, R. P. & Veje, C. *Phys. Rev. Lett.* **82**, 5241–5244 (1999).
- Edwards, S. F. in *Granular Matter: an Interdisciplinary Approach* (ed. Mehta, A.) 121–140 (Springer, New York, 1994).
- Mehta, A. & Edwards, S. F. *Physica A* **157**, 1091–1097 (1989).
- Nicodemi, M., Coniglio, A. & Herrmann, H. J. *Phys. Rev. E* **55**, 3962–3969 (1997).
- Nowak, E. R., Knight, J. B., Ben-Naim, E., Jaeger, H. M. & Nagel, S. R. *Phys. Rev. E* **57**, 1971–1982 (1998).
- Knight, J. B., Fandrich, C. G., Lau, C. N., Jaeger, H. M. & Nagel, S. R. *Phys. Rev. E* **51**, 3957–3963 (1995).

Daedalus

Perfect perforations

Birds have feathery porous wings, and engineers are pondering the reason, with a view to applying it to aircraft wings. One idea is to riddle aircraft wings with tiny holes. Any solid wing has a boundary layer of air. A porous wing can remove this layer by suction, or displace it by blowing. To optimize these effects, says Daedalus, we need a wing that is dense with little holes, but not at 90° to the wing surface. The wing should suck at the front, with each inlet hole pointing forwards to accept air from ahead. It should blow at the rear, with exit holes angled steeply to eject air backwards. The two effects would counterbalance each other, but need not cancel out. The ideal combination would optimize the wing as a lifting device of low drag.

Modern passenger aircraft need to provide cabin air for their passengers. The wings might draw on the plane's air budget, or alternatively might help to supply it; any supply from or drain to the wings must be taken into account. Even so, Daedalus reckons that porous wings must have good physics behind them, or rather beneath them. Nature knows her business.

Daedalus does not intend to power his aircraft from the air extracted ahead and ejected behind by its porous wings, though this should make a useful contribution. His idea is simply to make artificial porous wings as efficient as possible, and with the lowest feasible drag.

So DREADCO engineers are flying blown and sucked wings in a wind tunnel, and comparing the results with those obtained with naturally feathered wings. Both natural and artificially porous wings should have lower drag and higher lift than the standard solid variety. Further, sucking in air at the front while ejecting it at the back should improve the wings both as lifters and thrusters.

One problem will be ice, which at high altitudes grows on even the best wings. But conventional engines are only about 25% efficient, so 75% of their energy is wasted as heat. A distributed cooling system would warm the wings, helping to keep ice away. It could heat the rear-ejected air as well, increasing the thrust of blown wings by a sort of afterburner effect. A plane with cylinder engines could even release its exhaust gases through the rear-facing holes in a blown wing. This would neatly capture energy that would otherwise be lost, and raise the thrust of the wings. Sadly, it would disfigure the usual tasteless painted-on colour scheme.

David Jones

Obituary

Franco Rasetti (1901–2001)

When Franco Rasetti died on 5 December 2001, the scientific world lost one of its most prolific generalists. Rasetti was a member, with Enrico Fermi, of Italy's celebrated 'via Panisperna' research group in the 1930s. He was an acknowledged authority not only in spectroscopy, nuclear physics and cosmic radiation, but also in palaeontology and botany. He made valuable contributions in optics and entomology; published scores of papers; collected, classified and named hundreds of palaeontological specimens; and his book on Alpine flora has sold tens of thousands of copies.

Rasetti was born in Pozzuolo Umbro, Italy. He inherited a flair for natural science — his father was a professor of agricultural entomology and botany, and his mother was an accomplished artist, specializing in entomology. They educated the young Franco themselves. At just seventeen, he published his first paper, on the insects of Pisa and Lucca.

When the family moved to Pisa, Rasetti met a young Roman called Enrico Fermi. Fermi encouraged Rasetti to study physics rather than entomology and, on long walks, the two engaged in furious debates about scientific disciplines. In the end, Fermi won Rasetti over, and the two enrolled in the sparsely attended postwar physics course at the Pisa Institute. Rasetti then moved to Florence where he completed his PhD.

The friendship between the two matured into a fruitful professional collaboration. Rasetti was the master of experimental work, whereas Fermi had a more theoretical bent. By 1927 their reputation had grown such that Orso Mario Corbino invited them to join him at the Physics Institute of Rome. With Edoardo Amaldi, Emilio Segrè and Oscar D'Agostino, they became the famous via Panisperna group — named after the address of the institute — which was to revolutionize Italian, and indeed international, physics.

Inspired by Rasetti and Fermi, the group developed research programmes, enrolled graduate students and visited other laboratories. They imported experimental techniques — and invented and exported their own. In particular they started investigating the transmutation of elements by neutron bombardment. They quickly discovered sixty new radioactive nuclei and used slow neutrons to induce the fission of uranium. This work led to the patent of the chain-reaction process — of which Rasetti was the last surviving



Physicist and palaeontologist

holder — and eventually to the development of nuclear reactors and a multi-billion-dollar industry.

In the midst of this activity, Rasetti managed to spend some time at the California Institute of Technology in Pasadena. There, he cobbled together equipment to investigate the newly discovered Raman effect — the imprint on scattered light of the vibrational levels of a medium's molecules. His studies in 1929 of the effect in hydrogen, oxygen and nitrogen were hailed as the year's outstanding work in spectroscopy. His famous spectrum of the Raman effect in nitrogen has been reproduced in many texts.

As the storm clouds of the Second World War gathered, research conditions in Italy became difficult. In 1938, Segrè and Fermi left for the United States; Rasetti was courted by Université Laval in Québec, Canada, as well as several larger US institutions. He chose Laval to avoid becoming involved in war-related research — unlike his friend Fermi, who became a central figure in the Allied project to develop the nuclear bomb. Rasetti was not a pacifist: his objections to the war were not moral or political, rather, he simply considered war to be foolish and did not wish to be involved.

On his arrival at Laval in 1939, Rasetti was given the daunting task of creating a physics department from scratch. But within eight years a solid physics programme was turning out dozens of honours students, and research was burgeoning in several fields. It was at this time that I met Rasetti, and became his

last graduate student at Laval. He had developed equipment and techniques for studying high-energy cosmic rays, and had measured the half-life of the recently discovered particle, the ρ meson. Just as in Rome, his inventions and techniques were duplicated in many laboratories.

In Canada, Rasetti continued his life-long habits of walking and mountain-climbing. An interest in Canadian geology led him into palaeontology and then to trilobites — fossilized marine creatures of the Cambrian era. Rasetti came to be a major influence on the palaeontological scene. He sought out trilobite fossils everywhere, collecting thousands of specimens, which he described, classified and in many cases named. He developed techniques for preserving and photographing them for journals, and his collection rivalled that of Washington's Smithsonian Institution. Just as he had been honoured for his work on nuclear physics and Raman spectroscopy, Rasetti received numerous awards for his contributions to palaeontology; for example, in 1952 he won the Charles Doolittle Walcott Pre-Cambrian Research Medal, which is awarded every five years by the US National Academy of Sciences.

At the end of the Second World War, Rasetti finally moved to the United States, leaving Laval for Johns Hopkins University. There his research centred on radioactivity, particularly the gamma-ray spectra and decay schemes of thorium, ionium and radon. In 1949 he married Marie Madeline Hennin, a Belgian he had previously met in Québec. The couple retired to Waremmé, Belgium, in 1977, and it was there that Franco Rasetti died, aged 100.

Today, we hear a great deal about scientific fraud, and commissions and committees on scientific ethics abound. For Rasetti, scientific honesty was axiomatic and automatic. We speak of research teams and networks of research groups, but Rasetti considered that the autonomy and independence of the individual scientist were paramount. We hear of national scientific strategies and policies, but Franco Rasetti resisted politics, believing that the fruits of science were for all humanity.

His work and writing are noted for the elegance, simplicity and beauty which he considered to be the innate characteristics of science. Several books and biographies attest to this, most recently the biography by Danielle Ouellet and René Bureau: *Franco Rasetti — physicien et naturaliste* (Guérin, 2000).

Larkin Kerwin

Larkin Kerwin is Professor Emeritus at Université Laval, Québec G1K 7P4, Canada.

AJP

Carbon nanothermometer containing gallium

Gallium's macroscopic properties are retained on a miniature scale in this nanodevice.

Many applications have been found for carbon nanotubes^{1–8}, and we can now add a role as a 'nanothermometer' to this list. We describe how the height of a continuous, unidimensional column of liquid gallium inside a carbon nanotube (up to about 10 micrometres long and about 75 nanometres in diameter) varies linearly and reproducibly in the temperature range 50–500 °C, with an expansion coefficient that is the same as for gallium in the macroscopic state. We chose gallium as our thermal indicator because it has one of the greatest liquid ranges of any metal (29.78–2,403 °C) and a low vapour pressure even at high temperatures⁹. This nanothermometer should be suitable for use in a wide variety of microenvironments.

We investigated the behaviour of gallium inside carbon nanotubes (diameter, 40–150 nm) using a microscope equipped with a Gatan holder and twin heating system. Figure 1a–c shows a gallium column of diameter 75 nm and a continuous length of up to 7,560 nm. When the column temperature is increased or decreased in the range 50–500 °C, the gallium level rises or falls consistently.

The variation with temperature of the height of the gallium meniscus in its carbon nanotube is plotted in Fig. 1d, using the level at 58 °C as a reference. Changes in the length and diameter of the carbon nanotube itself can be disregarded because of the minute linear expansion coefficient¹⁰ of graphite (about -1×10^{-6} per °C at 20–500 °C), so the height of the gallium column is determined by its volume at that temperature.

The volumetric change of liquid gallium in the macroscopic state upon heating is described by $v_t = v_0(1 + \alpha\Delta t)$, where v_t and v_0 are the volumes at temperatures t and t_0 respectively, $\Delta t = t - t_0$, and α is the volumetric-expansion coefficient (0.1015×10^{-3} per °C at 30–977 °C, derived from measurements of density against temperature⁹). Our results indicate that this description also applies to nanoquantities of one-dimensional liquid gallium: calculations on the basis of its change in volume with temperature give a value for α of $0.095 \pm 0.006 \times 10^{-3}$ per °C, which is comparable to that for macroscopic liquid gallium. In this respect, the expansion coefficient differs from another basic thermal property, the melting point, which is greatly influenced by the surface effect¹¹.

Because the gallium meniscus level in a carbon nanotube moves linearly and reproducibly with temperature in the range 50–500 °C, it meets the requirements of a filled-system thermometer in this range. For such a nanothermometer, the temperature

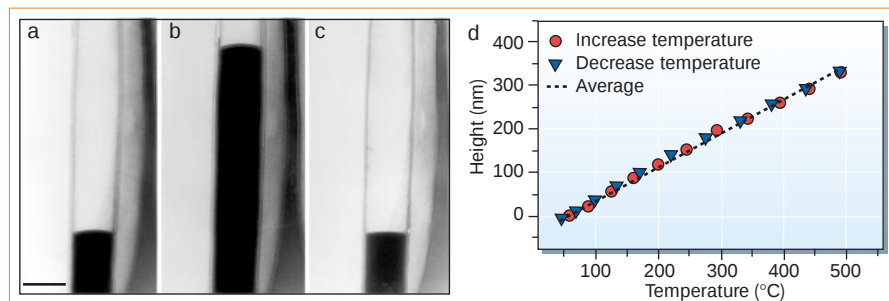


Figure 1 Expansion of gallium inside a carbon nanotube with increasing temperature. **a–c**, Changing level of the gallium meniscus at 58 °C (**a**), 490 °C (**b**) and 45 °C (**c**); scale bar, 75 nm. **d**, Height of the gallium meniscus plotted against temperature, measured in steps of 30–50 °C; results are averaged (green curve) from closely similar measurements obtained during heating (red) and cooling (blue). The nanothermometer was synthesized in a vertical radiofrequency furnace (which differs from a one-step arc-discharge method). A homogeneous mixture of Ga_2O_3 and pure, amorphous, active carbon (weight ratio, 7.8:1) was reacted in an open carbon crucible under a flow of pure N_2 gas: at 1,360 °C, the reaction $\text{Ga}_2\text{O}_3(\text{solid}) + 2\text{C}(\text{solid}) \rightarrow \text{Ga}_2\text{O}(\text{vapour}) + 2\text{CO}(\text{vapour})$ occurs. However, on the inner surface of a pure graphite outlet pipe at the top of the furnace, the temperature is lower (around 800 °C), causing the reaction $\text{Ga}_2\text{O}(\text{vapour}) + 3\text{CO}(\text{vapour}) \rightarrow 2\text{Ga}(\text{liquid}) + \text{C}(\text{solid}) + 2\text{CO}_2(\text{vapour})$ to occur, during which the 'nanothermometers' are created.

can be measured as $t = 58 + \Delta h/0.753$, where Δh (in nm) is the difference in the height of the gallium column at t °C and 58 °C.

It should be feasible to read the temperature recorded by the nanothermometer *in situ* with the help of a scanning electron microscope, given that the walls of the carbon nanotube and the space inside, as well as the gallium level, can be clearly seen even using an instrument operated at 10 keV. Because it has a measuring range of 50–500 °C, our nanothermometer will extend temperature measurement beyond the 4–80 K range attainable by resistance micrometre-sized cryogenic thermometers¹². It is easy to use, as the gallium meniscus is almost perpendicular to the inner surface of the carbon nanotube and the liquid column is continuous and long (up to 10 μm). The potential application proposed here follows on from several studies based on the discovery that carbon nanotubes can be filled with metal¹³.

Psychobiology

Speech sounds learned by sleeping newborns

It is not yet clear whether humans are able to learn while they are sleeping^{1,2}. Here we show that full-term human newborns can be taught to discriminate between similar vowel sounds when they are fast asleep. It is possible that such sleep training soon after birth could find application in clinical or educational situations^{3,4}.

We used mismatch negativity (MMN) to determine the ability of human newborns to detect a change in speech sounds.

Yihua Gao, Yoshio Bando

Advanced Materials Laboratory and Nanomaterials Laboratory, National Institute for Materials Science, Namiki 1-1, Tsukuba, Ibaraki 305-0044, Japan
e-mail: bando.yoshio@nims.go.jp

1. Tans, S. J., Verschueren, A. R. M. & Dekker, C. *Nature* **393**, 49–52 (1998).
2. Dai, H. J., Hafner, J. H., Rinzler, A. G., Colbert, D. T. & Smalley, R. E. *Nature* **384**, 147–150 (1996).
3. Tang, Z. K. *et al. Science* **292**, 2462–2465 (2001).
4. Poncharal, P., Wang, Z. L., Ugarte, D. & de Heer, W. A. *Science* **283**, 1513–1516 (1999).
5. Treacy, M. M. J., Ebbesen, T. W. & Gibson, J. M. *Nature* **381**, 678–680 (1996).
6. Kim, P. & Lieber, C. M. *Science* **286**, 2148–2150 (1999).
7. Kong, J. *et al. Science* **287**, 622–625 (2000).
8. Dillon, A. C. *et al. Nature* **386**, 377–379 (1997).
9. Lide, D. R. *CRC Handbook of Chemistry and Physics* 71st edn (CRC, Ohio, 1990–91).
10. Gray, D. E. *et al. American Institute of Physics Handbook* 3rd edn (McGraw-Hill, New York, 1972).
11. Wu, Y. Y. & Yang, P. D. *Adv. Mater.* **13**, 520–523 (2001).
12. Chantal, O., Bagueard, B., Bethoux, O. & Chabaud, B. *Rev. Sci. Instrum.* **68**, 2442–2446 (1997).
13. Ajayan, P. M. & Iijima, S. *Nature* **361**, 333–334 (1993).

Competing financial interests: declared none.

MMN is an attention-independent electrophysiological brain response which is elicited by infrequent discriminable changes in auditory stimuli⁵. MMN can be observed in young infants throughout all sleep stages as well as when they are awake^{6–8}.

We used MMN recordings to analyse the responses of three groups of full-term newborns (experimental group, $n = 15$, age 1–7 days; two control groups, $n = 15$ in both, age 2–7 days) to the vowel sounds /y/ ('standard', $p = 0.8$), /i/, and /y/i/ ('deviants', $p = 0.1$ for each) while they were asleep (for details of stimuli, see ref. 6). All groups underwent identical MMN recording sessions in the evening (session

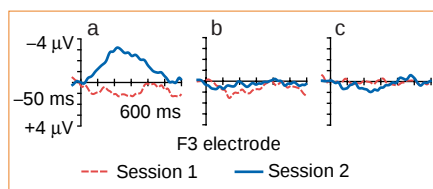


Figure 1 Mismatch negativity (MMN) results from two recording sessions, one in the evening and one the next morning, of responses to deviant speech sounds (/y/) in three groups (one experimental and two control groups) of human newborns. **a**, Experimental group, which underwent nocturnal training. The significant difference between results from the pre- and post-training sessions in this group suggests that sleeping newborns are better able to discriminate the speech sounds /y/ and /y/ after training. **b**, **c**, Control group 1 (**b**), which was not exposed to any training, and control group 2 (**c**), which was trained to discriminate between other vowel sounds, show no difference between the two sessions. The electro-encephalographic output was averaged and digitally filtered (1–30 Hz) at two frontal and two central scalp sites as well as two eye electrodes (for exclusion criteria and other details, see ref. 6). Mean MMN amplitude was calculated relative to a 60-ms period centred on the largest peak between 150 and 400 ms from stimulus onset in all except the eye electrodes. The presence of MMN was tested by using two-tailed *t*-tests for dependent samples.

1) and the next morning (session 2); the experimental group also had a session the following evening (session 3). Control group 1 received no training during the night, but the experimental group and control group 2 were exposed to nocturnal auditory training during sleep between sessions 1 and 2. Each MMN recording session lasted for just under 1 hour; in the experimental group, the training session was carried out over 2.5–5.0 hours. The experimental group and control group 2 were treated identically in sessions 1 and 2, except that during training the second control group was presented with /a/ and /e/ sounds instead of /y/ and /i/.

Throughout all sessions in all groups, a randomized sequence consisting of standards and deviants was presented online. In addition to pre-training stimuli with the experimental group, during session 2 we presented stimuli of different pitches that were otherwise identical: this was to determine whether the infants were simply learning acoustical discrimination or whether their increased discrimination ability was related to speech sounds.

The results from both pre- and post-training sessions showed that MMN responses to the acoustically 'easier' /i/ deviant stimulus ($t(14)$ varying between 2.5–5.8) were statistically significant in all groups. In session 1, the MMN amplitude was significantly smaller for the deviant /y/i/ than for /i/ in all groups, but no differences were found between the groups ($F(1,84)$, 8.62, $P < 0.004$) in a three-way ANOVA analysis (group X , stimulus X , electrode (F3, F4, C3, C4)). The difference between sessions 1 and 2 in all groups and

for both deviants was separately tested by using a two-way ANOVA (session X , electrode). In both control groups, no statistical difference in MMN amplitude was found between sessions 1 and 2.

Our main finding was that experimental subjects learned to discriminate both deviants from the standard after training (Fig. 1). In this group, the MMN elicited by /y/i/ was not significant in session 1 (mean, -0.28 ; s.d. 2.19) but it was in session 2 (mean, -3.76 ; s.d. 2.04). The difference between sessions 1 and 2 was $F(1,28)$, 16.58, $P < 0.0003$. Moreover, the MMN response for the /i/ deviant increased strikingly in amplitude after training ($F(1,28)$, 27.57, $P < 0.00001$). There was no significant reduction in MMN amplitude between sessions 2 and 3 for either deviant in the experimental group.

The experimental group learned not only to discriminate between the stimuli to which they were exposed during training, but also between stimuli (not presented to them during training) that had different F0 values (pitch) but were otherwise identical. The differences obtained in session 2 in response to the trained and non-trained stimuli were tested by using two-way ANOVAs (deviant (trained and untrained), X electrode): no significant deviant effect was found for either deviant. Moreover, the experimental group exhibited a significant MMN response to both deviants in session 3, indicating that the training effect lasted for some time. In separate two-way ANOVAs (session X , electrode), no significant session effect was found for either deviant.

We have shown that newborns can assimilate auditory information while they are sleeping, suggesting that this route to learning may be more efficient in neonates than it is generally thought to be in adults.

M. Cheour*, **O. Martynova***,
R. Näätänen†‡, **R. Erkkola||**, **M. Sillanpää¶**,
P. Kero#, **A. Raz☆**, **M.-L. Kaipio‡**,
J. Hiltunen*, **O. Aaltonen****, **J. Savela****,
H. Hämäläinen*

*Language and the Developing Brain Laboratory, Centre for Cognitive Neuroscience, and Department of Psychology, and **Department of Phonetics, University of Turku, 20100 Turku, Finland
e-mail: marie.cheour@utu.fi

†BioMag Laboratory, Helsinki University Central Hospital, 00029 Helsinki, Finland

‡Cognitive Brain Research Unit, Department of Psychology, and §Helsinki Brain Research Centre, University of Helsinki, 00014 Helsinki, Finland

¶Obstetrics and Gynaecology, ||Child Neurology, and #Pediatrics, Turku University Central Hospital, 20100 Turku, Finland

☆Sackler Institute for Developmental Psychobiology, Department of Psychiatry, Weill Medical College of Cornell University, New York, New York 10021, USA

1. Dave, A. S., Yu, A. C. & Margoliash, D. *Science* **282**, 2250–2254 (1998).
2. van Hateren, C. F., Boekkooi, P. F., Jongsma, H. W. & Nijhuis, J. G. *Lancet* **356**, 1169–1170 (2000).
3. Tallal, P. P., Miller, S. L., Bedi, G. G., Byma, G. G. & Wang, X. X. *Science* **271**, 77–81 (1996).
4. Kraus, N. et al. *Science* **273**, 971–973 (1996).
5. Näätänen, R. et al. *Nature* **358**, 432–434 (1997).
6. Cheour, M. et al. *Int. J. Psychophys.* **29**, 217–226 (1998).
7. Cheour, M. et al. *Nature Neurosci.* **1**, 351–353 (1998).
8. Cheour, M., Leppänen, P. H. & Kraus, N. *Clin. Neurophysiol.* **117**, 4–16 (2000).

Competing financial interests: declared none.

Microstructures

Spin-engineering magnetic media

The explosion in demand for increased data-storage density is driving the exploration of new magnetic media. Here we describe a new type of magnetic medium in which the spin configurations are engineered in chemically homogeneous magnetic films: regularly arranged in-plane and out-of-plane spin configurations are defined by altering the magnetic anisotropy. These spin-engineered media not only maintain the surface planarity but also the homogeneity of the magnetic materials, and our method is likely to find immediate application on account of its simplicity and ease of integration.

Two approaches are being used in the search for ultrahigh-density magnetic recording media. One is based on a conventional, continuous film medium, in which the storage density is increased by stabilizing the nanoscale magnetic domains in order to resist the thermal self-erasure effect¹. The other depends on the patterning of magnetic films into magnetically isolated dots², with each dot containing two discrete magnetized states with equal but opposite magnetic moments. Each dot is thus able to store one bit of information.

We used a modulated single/polycrystalline substrate surface to modify locally the magnetic anisotropy in subsequently deposited magnetic films, which induces the desired artificial magnetic structure. Selective epitaxial growth introduces an alternation between single-crystal and polycrystalline structures in the film, according to the substrate patterning (Fig. 1a). Epitaxial Ni/Cu(001) films of appropriate thickness show perpendicular magnetization, which is due to the magneto-elastic interaction induced by the Ni/Cu(001) interface^{3,4}. In contrast, the magnetization of polycrystalline nickel lies in the film plane because of the dominant demagnetizing field. The substrates used are GaAs(001) and two types of pattern were chosen: wire and dot array.

The arrays were obtained by lithography

and the subsequent lifting off of 1-nm-thick nickel (this thickness was chosen for simplicity — in principle, half a monolayer is sufficient for selective epitaxial growth). The ultra-thin nickel patterns were then oxidized in air to become ultra-thin nickel oxide templates. The patterned substrates were annealed to clean the surface before growing Cu(5 nm)/Ni(5 nm)/Cu(70 nm)/Co(1.8 nm) films at room temperature. The Cu/Co films are used as an underlayer⁵ to promote epitaxial nickel growth on the substrate, with 5-nm-thick nickel displaying a stable perpendicular magnetization^{3,6}. Films grown on the NiO templates are polycrystalline, whereas those grown directly onto the GaAs surface are single-crystal films. We examined the sample surface by atomic-force microscopy, which yielded roughness parameters in a similar range to that of epitaxial continuous Cu/Ni/Cu films⁴.

The modulated magnetic structures were confirmed by magnetic-force microscopy (MFM) and magneto-optic Kerr-effect (MOKE) magnetometry. Figure 1b shows the MFM images of the dot and wire samples in the remanent state after saturation with a perpendicular field. The bright stripes in the images correspond to the out-of-plane nickel magnetization, which has a strong magnetic signal compared with that

of the in-plane magnetization.

Figure 1c shows MOKE measurements in the field-perpendicular-to-film (polar) geometry. The perfectly square hysteresis loop obtained from the unpatterned epitaxial film indicates that only perpendicular magnetization exists (Fig. 1c, left). For the wire sample, however, the sharp switch at low field originates from the regions of perpendicular magnetization (Fig. 1c, right), whereas the gradual saturation in high field is caused by the presence of the in-plane magnetized regions.

The spin-engineering of this magnetic medium, which can be seen as the integration of patterned features into chemically homogeneous films, opens up new avenues for controlling the spin structure of magnetic materials. Our simulation results show that the patterned bit resolution can be smaller than 30 nm. Such modulated structures can be used as planar patterned magnetic media⁷ without breaking the homogeneity of the magnetic film, which is essential for avoiding effects caused by reduced Curie temperature. Furthermore, the anisotropy-constrained magnetic walls are static in nature, which provides a reproducible switching mechanism, unlike that of the natural domain wall. Such a system could also be useful for studying magnetic dipolar and exchange

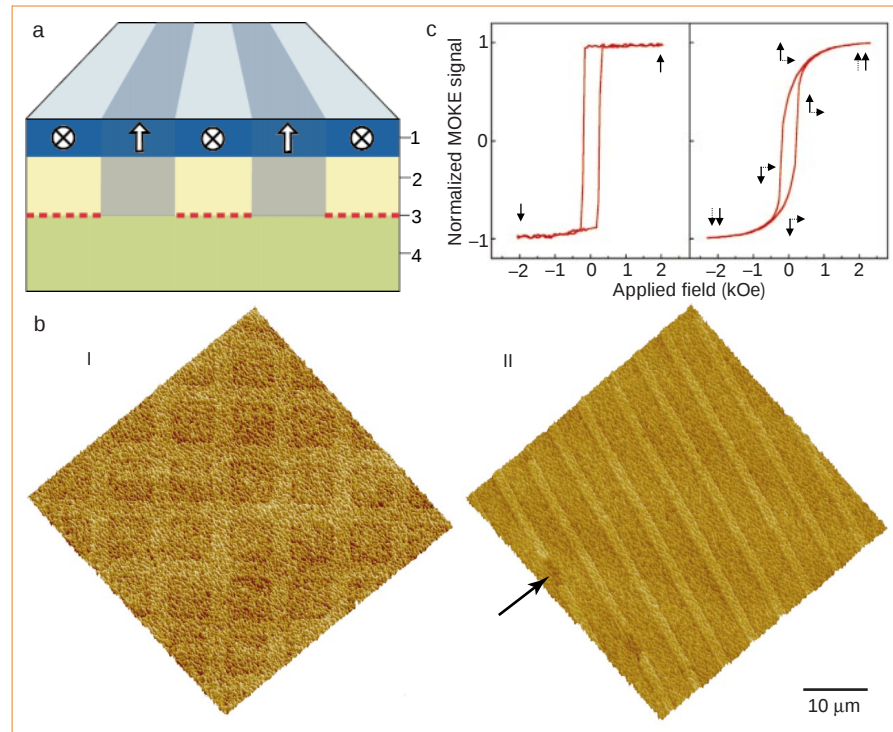


Figure 1 Selective epitaxy in spin engineering and magnetic characterization. **a**, Selective epitaxial growth of Cu/Ni/Cu/Co structure. Layer 1, magnetic layer of nickel; symbols: crosses, magnetization in the plane of the film; arrows: magnetization perpendicular to the film. Layer 2, Cu/Co underlayer. Ultra-thin NiO was used as the embedded template (layer 3) on a GaAs(001) substrate (layer 4). The wire arrays are 2 μm wide and separated by 4 μm; the square dots are 7 × 7 μm² and are separated by 3 μm. **b**, Magnetic-force microscopy (MFM) images of the dot (I) and wire (II) samples in zero field after perpendicular field saturation. Arrow indicates a switch of the perpendicular Ni induced by the external stray field of the MFM tip. **c**, Polar magneto-optic Kerr-effect (MOKE) hysteresis loops for the unpatterned reference sample (left panel) and the wire sample (right panel). Solid and dotted arrows indicate the magnetization orientation in the epitaxial and polycrystalline regions, respectively, of the Ni film.

interactions, for example, and domain-wall resistance in highly controlled geometric systems.

S. P. Li*, W. S. Lew*, J. A. C. Bland*, L. Lopez-Diaz*, M. Natali†, C. A. F. Vaz*, Y. Chen†

*Cavendish Laboratory, University of Cambridge, Madingley Road, Cambridge CB3 0HE, UK
e-mail: jacb1@phy.cam.ac.uk

†Laboratoire de Microstructures et de Microélectronique (CNRS), 196 Avenue Henri Ravera, 92225 Bagneux Cedex, France

1. Krusin-Elbaum, L., Shibauchi, T., Argyle, B., Gignac, L. & Weller, D. *Nature* **410**, 444–446 (2001).
2. Chou, S. Y. *Proc. IEEE* **85**, 652–671 (1997).
3. O'Brien, W. L. & Tonner, B. P. *Phys. Rev. B* **49**, 15370–15373 (1994).
4. Hope, S. *et al. Phys. Rev. B* **55**, 11422–11431 (1997).
5. Lew, W. S. *et al. J. Appl. Phys.* **87**, 5947–5949 (2000).
6. Bochi, G., Ballentine, C. A., Ingelfield, H. E., Thompson, C. V. & O'Handley, R. C. *Phys. Rev. B* **53**, 1729–1732 (1996).
7. Chappert, C. *et al. Science* **280**, 1919–1922 (1998).

Competing financial interests: declared none.

COMMUNICATIONS ARISING

Ecology

Is coral bleaching really adaptive?

From an experiment in which corals are transplanted between two depths on a Panamanian coral reef, Baker¹ infers that bleaching may sometimes help reef corals to survive environmental change. Although Baker's results hint at further mechanisms by which reef-building corals may acclimatize to changing light conditions, we do not consider that the evidence supports his inference¹.

Baker's study attempts to test the 'adaptive bleaching hypothesis'² (ABH), in which stressed corals first lose their dinoflagellate symbionts (bleach) and then regain a new mix of symbionts that are better suited to the imposed stress regime³. Bleaching in response to increased, and not decreased, irradiance is well known⁴, and Baker's corals suffered more bleaching when transferred to the shallow site. However, their mortality was lower and their mix of symbiont genotypes was altered, unlike those corals that were transplanted to the deeper site, leading Baker to conclude that the higher mortality of the latter corals was due to their failure to bleach and hence to vary their symbiont genotypes.

We are concerned not only that Baker uses undefined stresses, which do not specifically include temperature (thought to be the primary cause of mass coral bleaching⁵), but also that the corals in his treatments recover under very different conditions. The corals classified as "chronically stressed" recovered at the relatively low-light, deep-water site (20–23 m),

where the only two species that showed increased mortality at depth (*Diploria strigosa* and *Acropora cervicornis*) were very rare. Baker's "acutely stressed" corals, however, recovered under the higher light levels of a shallow-water site (2–4 m). From this experimental design, we cannot unequivocally conclude that the improved survival of the acutely stressed corals was due to their adoption of a new mix of dinoflagellates after bleaching, or to improved recovery conditions at the shallow site. As light energy is critical to the survival of reef-building corals⁶, stressed corals might be expected to survive better when transplanted to a more sunlit site and less well after transfer to deep water, irrespective of bleaching.

The ABH assumes that bleached corals favour new host–symbiont associations that optimize survival, necessitating rapid evolutionary adaptation (that is, genetic change) by populations of reef-building corals and their symbionts³. Although Baker claims that bleaching offers an ecological opportunity for reef corals to rid themselves rapidly of suboptimal algae and to acquire new partners¹, he relies on a molecular technique that is unable to distinguish newly invading genotypes from other rare genotypes that are already present in the host and which simply increase in proportion after conditions change. The latter is a phenotypic change (acclimatization) and, as such, is restricted in its provision of new genetic combinations for evolution.

We consider that the evidence in favour of the ABH remains scant in the absence of observations that the genotypes of symbionts in corals become more thermally robust during and after mass bleaching. Baker's finding that corals adopt a different mix of symbiont genotypes when moved from one light environment to another is an interesting addition to the well-known acclimatory responses of corals and their symbionts to changes in light quality and quantity⁷, but we cannot conclude that bleaching favours new host–symbiont combinations that guard populations of corals against rising sea temperature.

Ove Hoegh-Guldberg, Ross J. Jones, Selina Ward, William K. Loh

Centre for Marine Studies, University of Queensland, Queensland 4072, Australia
e-mail: oveh@uq.edu.au

Baker replies — Hoegh-Guldberg *et al.* suggest that corals that were transplanted downwards died more frequently than those transplanted upwards because they were deprived of critical sunlight energy at depth. My argument went a step further by explaining why this energy is so critical for these transplanted colonies.

Because corals that were transplanted downwards did not bleach in response to reduced irradiance, they failed to exchange their 'high-light' algal symbionts for the more suitable 'low-light' algae that were already found in the deep-water colonies at this site (and/or at other sites nearby). As a result, they contained inappropriate algae for their new environment, which led to chronic stress and eventual mortality.

In contrast, corals that were transplanted upwards experienced severe bleaching as a result of increased irradiance. Consequently, suboptimal low-light algae were removed, allowing high-light algae to become dominant in the newly vacant hosts. Such corals survived well as a result, despite their initial bleaching. This explanation is particularly powerful because it unifies coral bleaching, symbiont change and host mortality.

Hoegh-Guldberg *et al.* suggest that my findings fail to support the ABH because they do not provide evidence of 'new' symbionts in transplanted corals. The ABH is not limited to this constraint. Regardless of the origin of replacement symbionts (which, as I pointed out, may "colonize" and/or "proliferate inside" hosts) or the proximate environmental causes of bleaching (for example, light or temperature), if bleached reef corals change the composition of their symbiont communities faster than unbleached corals, and if more rapid symbiont change proves beneficial, then bleaching has adaptive value. Even if adult colonies are unable to form symbioses with unusual or new algae (which is unlikely, given the recent discovery of some scleractinian coral colonies containing symbionts that are usually found in foraminifera¹), cryptic populations of diverse symbionts may still occur at low abundance in many coral hosts².

There is no field evidence that symbiont genotypes change after bleaching events because the necessary molecular investigations have not yet been undertaken. Despite this, one of the best available long-term data sets on mass coral bleaching and mortality reveals that far fewer corals in the far-eastern Pacific Ocean died after the 1997–98 El Niño event (0–26%) than after the 1982–83 El Niño event (52–97%; ref. 3), even though the magnitude and duration of sea-surface temperature anomalies in the region in 1997–98 exceeded those of 1982–83 (ref. 4). These observations indicate that surviving reef corals may be more

resistant to recurrent thermal stress through having experienced earlier episodes of severe bleaching and mortality, as predicted by models of symbiont change⁵.

Furthermore, field experiments with bleached corals⁶ and laboratory studies of model invertebrate–algal symbioses⁷ support some of the assumptions of the ABH. We should not mistake an absence of evidence for evidence of absence, and instead need to document worldwide patterns of coral–algal associations and their response to mass-bleaching events. The real question is not whether coral–algal associations can adapt by recombining, but rather how, and over what timescales, they do so.

Although episodes of mass coral bleaching and mortality will occur in the future, my findings suggest that they may not recur with the frequency and severity predicted by some studies⁸. This should stimulate efforts to protect the remaining three-quarters of the world's coral-reef ecosystems⁹ by reducing the compounding effects of anthropogenic factors that are still under our influence.

Andrew C. Baker*

*Marine Conservation Programs, Wildlife Conservation Society, 2300 Southern Boulevard, Bronx, New York 10460, USA

†Center for Environmental Research and Conservation, Columbia University, New York 10027, USA

e-mail: abaker@wcs.org

- Rodriguez-Lanetty, M., Cha, H. R. & Song, J. I. *Proc. 9th Int. Coral Reef Symp.* (in the press).
- Goulet, T. L. & Coffroth, M. A. *Proc. 8th Int. Coral Reef Symp.* 2, 1331–1334 (1997).
- Glynn, P. W., Maté, J. L., Baker, A. C. & Calderón, M. O. *Bull. Mar. Sci.* 69, 79–109 (2001).
- Enfield, D. B. *Bull. Mar. Sci.* 69, 7–25 (2001).
- Ware, J. R., Fautin, D. G. & Buddemeier, R. W. *Ecol. Model.* 84, 199–214 (1996).
- Toller, W. W., Rowan, R. & Knowlton, N. *Biol. Bull. Mar. Biol. Lab. Woods Hole* 201, 360–373 (2001).
- Kinzie, R. A., Takayama, M., Santos, S. R. & Coffroth, M. A. *Biol. Bull. Mar. Biol. Lab. Woods Hole* 200, 51–58 (2001).
- Hoegh-Guldberg, O. *Mar. Freshwat. Res.* 50, 839–866 (1999).
- Wilkinson, C. (ed.) *Status of Coral Reefs of the World: 2000* (Austral. Inst. Mar. Sci., Townsville, Queensland, 2000).

errata

Seeing through the face of deception

I. Pavlidis, N. L. Eberhardt, J. A. Levine
Nature 415, 35 (2002)

It was not intended to convey the impression that this thermal-imaging technique is already suitable for mass security-screening purposes: indeed, the false-positive rate identified in this small study might preclude large-scale application.

Laterality in tool manufacture by crows

Gavin R. Hunt, Michael C. Corballis, Russell D. Gray
Nature 414, 707 (2001)

The tool held in the beak of the bird shown in Fig. 1 of this communication was wrongly described as a crochet tool, whereas it is a simple leaf-stem tool that happens to be hooked.

- Baker, A. C. *Nature* 411, 765–766 (2001).
- Buddemeier, R. W. & Fautin, D. G. *BioScience* 43, 320–326 (1993).
- Ware, J. R., Fautin, D. G. & Buddemeier, R. W. *Ecol. Model.* 84, 199–214 (1996).
- Jones, R. J. & Hoegh-Guldberg, O. *Plant Cell Envir.* 24, 89–100 (2001).
- Hoegh-Guldberg, O. *Mar. Freshwat. Res.* 50, 839–866 (1999).
- Muscattine, L. & Cernichiaro, E. *Biol. Bull. Mar. Biol. Lab. Woods Hole* 137, 506–523 (1969).
- Falkowski, P. G. & Dubinsky, Z. *Nature* 289, 172–174 (1981).

Slowdown of the meridional overturning circulation in the upper Pacific Ocean

Michael J. McPhaden* & Dongxiao Zhang*‡

* NOAA/Pacific Marine Environmental Laboratory, Seattle, Washington 98115, USA

‡ Joint Institute for the Study of the Atmosphere and Ocean, University of Washington, Seattle, Washington 98115, USA

Decadal temperature fluctuations in the Pacific Ocean have a significant effect on marine ecosystems and the climate of North America. The physical mechanisms responsible for these fluctuations are poorly understood. Some theories ascribe a central role to the wind-driven meridional overturning circulation between the tropical and subtropical oceans. Here we show, from observations over the past 50 years, that this overturning circulation has been slowing down since the 1970s, causing a decrease in upwelling of about 25% in an equatorial strip between 9° N and 9° S. This reduction in equatorial upwelling of relatively cool water, from 47×10^6 to $35 \times 10^6 \text{ m}^3 \text{ s}^{-1}$, is associated with a rise in equatorial sea surface temperatures of about 0.8 °C. Another effect of the slowing circulation is a reduction in the outgassing of CO₂ from the equatorial Pacific Ocean—at present the largest oceanic source of carbon dioxide to the atmosphere.

Recent analyses of historical data have revealed the signatures of decadal-timescale variations in the ocean–atmosphere system of the Pacific basin. These variations have been characterized in terms of cyclic oscillations over a broad band of frequencies with periods between 10 and 70 years¹. Alternatively, they have been described in terms of regime shifts—abrupt and coherent changes in climatological conditions affecting large areas of the Pacific basin—the most pronounced of which in the past 50 years occurred in 1976–77 (ref. 2). One representation of this decadal variability, the so-called Pacific Decadal Oscillation (PDO), has a spatial structure very similar to that of the El Niño/Southern Oscillation (ENSO), which fluctuates between warm (El Niño) and cold (La Niña) phases at shorter periods of 3–7 years (refs 3, 4). Like ENSO, decadal-timescale temperature fluctuations in the Pacific have substantial effects on marine ecosystems and the climate of North America^{5,6}. The dynamics of decadal fluctuations are much less well understood than those of ENSO, however, and there is considerable debate about the underlying physical processes that govern them, especially in the ocean⁷. Some theories invoke mid-latitude ocean–atmosphere interactions originating in the North Pacific Ocean⁸, while others point to the importance of ocean–atmosphere interactions spanning tropical and subtropical latitudes^{9–11}.

One of the most prominent theories for Pacific decadal climate variability¹¹ assumes that the meridional overturning circulation in the upper ocean acts as a conveyor belt for transporting water mass anomalies from the subtropics to the tropics. In this theory, air–sea interactions in the subtropics create surface water masses with unusually warm or cold temperatures that are subducted—that is, pushed downward—into the pycnocline (where the water density increases rapidly with depth). The temperature-tagged waters slowly drift equatorward below the surface for roughly 10 years, during which time they are isolated from further interaction with atmosphere. When these waters reach equatorial latitudes, they are upwelled to the surface where they are presumed to affect equatorial sea surface temperatures, air–sea heat exchange, and overlying atmospheric circulation in a manner similar to what happens during ENSO events. The upwelled water then flows back towards the subtropics in the surface layer to complete one circuit of the meridional circulation cell.

But recent evidence suggests that temperature anomalies subducted into the pycnocline at subtropical latitudes may not reach the Equator with any appreciable amplitude¹². Model studies have indicated that these anomalies can be either strongly

dissipated, disperse in the form of planetary-scale oceanic waves, or become obscured by wind-forced variations at low latitudes^{13–15}. Thus, a related hypothesis for the role of the meridional overturning circulation in decadal-timescale climate variability has been proposed: the circulation speeds up or slows down, transporting water masses with relatively constant properties at faster or slower rates towards the Equator¹⁶. A dynamically changing ocean circulation has significant implications for equatorial upwelling—which affects not only tropical Pacific sea surface temperatures, but also the exchange of CO₂ across the air–sea interface and the supply of nutrients to the biologically productive surface layer of the ocean.

To address the question of whether the meridional overturning circulation in the upper Pacific may be changing on decadal timescales, we have analysed hydrographic data between 20° S and 50° N for the period 1950–99 in the depth range of the upper pycnocline (typically 50–400 m in the tropical and subtropical Pacific). We have also computed surface-layer Ekman transports using multi-decade long wind products derived from empirical and atmospheric model-based analyses of available observations. Results of our analysis, and implications for climate variability, are described below.

Mean circulation in the pycnocline

To illustrate important features of the general circulation, and to set the context for discussion of decadal variations in pycnocline flows, Fig. 1a shows the 50-year mean geostrophic streamlines and horizontal velocity field on the 25.0 kg m^{-3} potential density surface which is located in the middle of the upper pycnocline. This density surface outcrops at the sea surface between about 30° and 40° N in the Northern Hemisphere and south of 20° S (except in the far eastern Pacific) in the Southern Hemisphere. Water subducted into the pycnocline along these outcrop lines can make its way to the Equator by both interior and western-boundary routes. Interior pathways are more circuitous in the Northern Hemisphere because of the presence of a potential vorticity ridge along 9°–10° N (associated with uplift of the pycnocline under the Intertropical Convergence Zone), which tends to block the flow of waters between higher latitudes and the Equator in the eastern and central Pacific¹⁷. Based on our estimates of geostrophic velocity, transit times to the Equator on this density surface are about 5–10 years, depending on details of a water parcel's pathway and where it is subducted. Comparable transit times and circulation pathways have

been inferred from numerical model simulations¹⁸, circulation theory¹⁹ and tracer distributions^{20,21}.

The flow pattern shown in Fig. 1a is similar to that on isopycnals between about 22 and 26 kg m⁻³ in the Northern Hemisphere and on isopycnals between 22.5 and 26.2 kg m⁻³ in the Southern Hemisphere. Waters less dense than 22.0 kg m⁻³ almost always fall within the surface mixed layer in the tropics, while interior flow on density surfaces deeper than 26.2 kg m⁻³ is weakly poleward near the Equator¹⁷. Thus, for the purpose of this analysis, we define interior pycnocline transports associated with the shallow meridional overturning circulation as occurring within the above range of density classes.

We calculate pycnocline transports by integrating equatorward flows along 9° N and 9° S from the eastern boundary to near the western boundary. We choose 9° N because it represents a choke point for meridional geostrophic transports in the interior ocean. No similar constriction is present south of the Equator, so we integrate along 9° S for hemispheric symmetry. To test for continuity with meridional transports from higher latitudes, we repeated these calculations at 15° N and 15° S, which approximate the boundaries separating the subtropical and tropical gyres in terms of the depth-integrated horizontal mass transport streamfunction²².

Decadal variations in pycnocline transports

We calculated meridional mass transports for 7–10 year periods in such a way as to span the 1976–77 regime shift, and to maximize utilization of the data that are not uniformly distributed in time. As expected from previous work¹⁷, transports are higher in the Southern Hemisphere than in the Northern Hemisphere where flow between the subtropics and tropics is more obstructed in the interior ocean (Fig. 2a). However, we find that these transports have

been non-stationary, with comparable decreases of 6–7 sverdrups (1 Sv = 10⁶ m³ s⁻¹) across both 9° N and 9° S since the 1970s.

The combined equatorward transport (or convergence) of pycnocline water across 9° N and 9° S drops from 27.0 ± 2.5 Sv before the regime shift in the 1970s to 14.0 ± 1.5 Sv in the 1990s (Table 1). For comparison, meridional transport convergences decrease from 38.0 ± 1.8 Sv in the 1970s to 29.5 ± 1.9 Sv in the 1990s across 15° N and 15° S in the same density classes. Larger absolute transports are expected at 15° N and 15° S because some of the interior ocean transport crossing the gyre boundaries is eventually shunted into equatorward-flowing western boundary currents. However, the downward trend is similar in magnitude to that observed across 9° N and 9° S, and is likewise the result of approximately equal decreases in Northern and Southern Hemisphere transports. Differences in the downward trend (8.5 Sv across 15° versus 13 Sv across 9°) may in part be due to the fact that some of the very lightest waters between 22 and 23 kg m⁻³ are formed equatorward of 15° N and 15° S. Decadal changes in western boundary current transports between 9° and 15° may also account for some of the difference, though (as will be shown below) changes in these transports of only a few sverdrups are not detectable with confidence from our analysis.

The period from the 1970s to the 1990s coincides with a rise in sea surface temperatures of about 0.8 °C in the equatorial Pacific⁴ (Fig. 2b). This temperature increase cannot be explained by an increase in surface heating. On the contrary, heat fluxes into the ocean have decreased over the past 50 years in the equatorial Pacific, primarily owing to an increase in cloudiness that has reduced the amount of sunlight reaching the surface²³. Thus, as occurs on ENSO timescales, changes in surface heat flux act to damp rather than to generate decadal changes in equatorial sea surface temperatures.

What accounts for the observed decadal decreases in interior

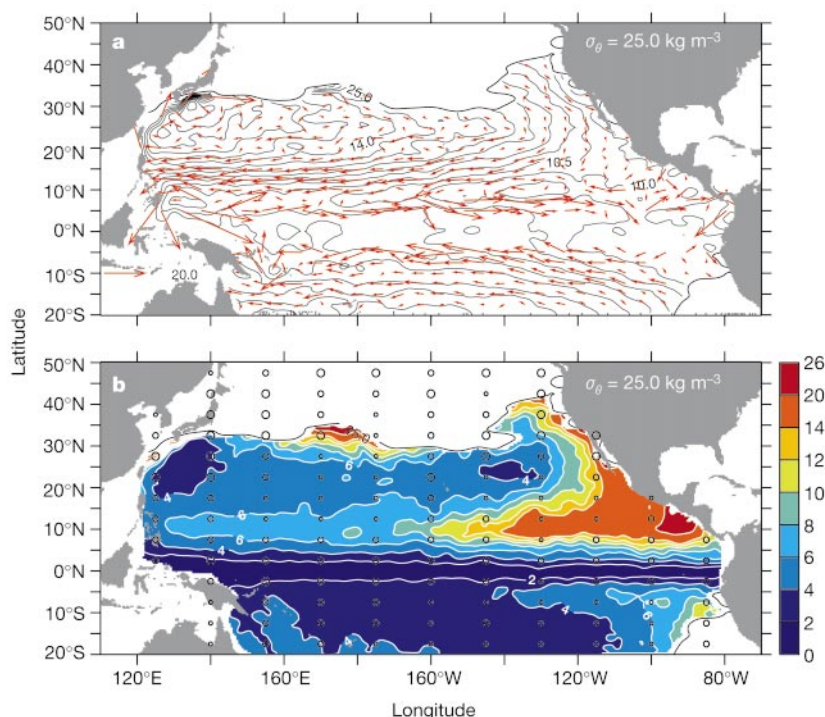


Figure 1 Mean circulation and potential vorticity averaged over 50 years (1950–99) in the upper pycnocline of the Pacific Ocean. **a**, Geostrophic streamlines relative to 900 dbar (in m⁻² s⁻²) and **b**, absolute value of potential vorticity (in 10⁻¹⁰ m⁻¹ s⁻¹) on the 25.0 kg m⁻³ potential density surface. Velocity vectors are overplotted on **a**. Distribution of hydrocasts down to 900 dbar is overplotted on **b**, with the size of the dots representing total number of casts in regions of 5° latitude by 15° longitude (smallest, 50–300; intermediate, 301–1,000; largest, 1,001 or more). Winter season outcrop lines are

drawn in both panels. The outcrop lines define locations where wintertime surface mixing penetrates deepest into the pycnocline, creating new water masses that are subsequently sequestered from the atmosphere as seasonal heating restratifies the upper ocean in spring and summer. Potential vorticity is defined as $f(\partial\rho/\partial z)/\rho_0$, where f is the Coriolis parameter, $\partial\rho/\partial z$ is the vertical density gradient, and ρ_0 is a constant reference density (10³ kg m⁻³). Water parcels conserve their potential vorticity in an ideal fluid.

ocean geostrophic transports, and are they related to the increases in sea surface temperature? Comparing the periods 1970–77 and 1990–99, we see that associated with the increase in sea surface temperature there is a decrease in the intensity of the easterly trade winds, as indicated by the prevalence of persistent westerly wind stress anomalies in the central and western Pacific (Fig. 3a). These features are the tropical manifestation of the PDO in its ‘high’ phase (warm tropics, cold central North Pacific)^{3,4}. Weaker surface easterlies in the 1990s are accompanied by increased surface wind convergence, atmospheric deep convection, and rainfall in the central equatorial Pacific, similar to what occurs during El Niño events²⁴.

The pycnocline normally slopes down to the west near the Equator, under the influence of easterly trade wind forcing. This pycnocline slope induces a geostrophically balanced meridional inflow towards the Equator in both hemispheres. The pattern of wind stress differences in Fig. 3a favours a relaxation of this pycnocline tilt, with anomalous deepening in the east and shoaling in the west. Consistent with this expectation, the depth of the pycnocline (as indicated by 25.0 kg m⁻³ surface) becomes shallower by about 10–30 m west of 160° W and deeper by about 10–30 m between 120°–150° W in the 1990s relative to the 1970s (Fig. 3b). As

a consequence, equatorward geostrophic transport is reduced in both hemispheres in the 1990s, in accordance with ocean circulation theory that predicts less interior pycnocline exchange for anomalous westerly winds near the Equator²⁵.

Decadal variations in surface layer transports

We expect that reduced meridional convergence in pycnocline transports should be associated with reduced poleward transport of water in the surface layer. To test this hypothesis, we computed meridional Ekman transports integrated from the eastern boundary to the western boundary of the Pacific for four different wind stress products that span at least two decades. Meridional Ekman transport per unit longitude is given by the expression $M_e = -\tau^x/\rho_0 f$, where τ^x is the zonal component of the wind stress, f is the Coriolis parameter (twice the local vertical component of the Earth’s angular velocity), and ρ_0 is a constant reference density (10³ kg m⁻³). These transports are confined approximately to the upper 50 m in the tropical oceans²⁶, and are to the right of the wind stress in the Northern Hemisphere and to the left of the wind stress in the Southern Hemisphere.

Most wind stress products individually show decreases in zonally integrated Ekman transports in the tropical Pacific associated with the broad-scale weakening of the trade winds between the 1970s and 1990s. However, to reduce the effects of uncertainties in individual products, and because no product is unambiguously superior, we computed ensemble means of Ekman transports and transport divergences from the four products for the same periods as for the meridional pycnocline transports in Fig. 2. Ensemble means and standard errors for the Ekman transport divergences across 9° N and 9° S, computed as $\int [M_e(9^\circ \text{N}) - M_e(9^\circ \text{S})] dx$, are 53.9 ± 5.3 Sv (1956–65), 58.4 ± 4.2 Sv (1970–77), 52.9 ± 4.5 Sv (1980–89) and 45.9 ± 5.5 Sv (1990–99). The Ekman transport divergence is highest during 1970–77 and lowest during 1990–99, decreasing by about 12 Sv over this time. This decrease is comparable to our estimated decrease in interior pycnocline transport convergence.

The difference between surface-layer transport divergence and pycnocline transport convergence across 9° N and 9° S should reflect flow in the western boundary currents (specifically, in the equatorward-flowing Mindanao Current across 9° N, and in the New Guinea Coastal Current and Undercurrent across 9° S). Inferred western boundary current transport convergences, estimated as the residual of a mass balance for each decadal period, are typically around 25 Sv (Table 1). Several more sverdrups at minimum should be added to these convergences, as western boundary currents (and the Mindanao Current in particular) supply upper pycnocline waters to feed much of the flow that exits the Pacific basin into the Indian Ocean through the Indonesian straits²⁷. It is noteworthy, though, that these western boundary transport convergences, whose magnitudes are broadly consistent with those from modelling and observational studies^{17,18}, do not seem to vary significantly on decadal timescales given the confidence limits of our estimates. Thus, if changes in western boundary current transports have

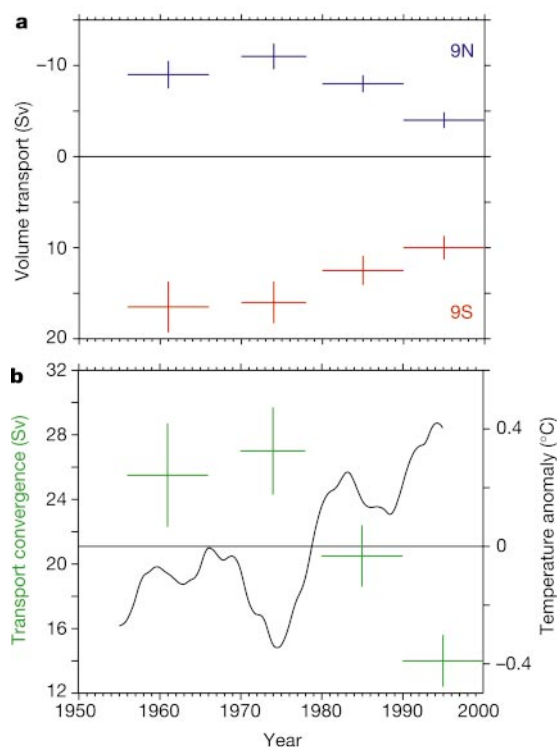


Figure 2 Meridional transports in the pycnocline and smoothed sea surface temperatures over the past 50 years. **a**, Mean zonally integrated meridional transports in the pycnocline relative to 900 dbar along 9° N and 9° S, computed for 1956–65, 1970–77, 1980–89 and 1990–99. Values are integrated in the Northern Hemisphere from the eastern boundary to 145° E in density classes between 22 and 26 kg m⁻³, and in the Southern Hemisphere from the eastern boundary to 160° E in density classes between 22.5 and 26.2 kg m⁻³. Transports are in units of sverdrups (1 Sv = 10⁶ m³ s⁻¹) which is the volumetric equivalent of mass for a constant reference density. Error bars are for one standard error. **b**, Mean meridional transport convergence (in Sv) in the pycnocline across 9° N and 9° S. Convergence is calculated as the difference between Southern Hemisphere minus Northern Hemisphere transports in **a**. Also plotted in **b** are areally averaged sea surface temperature anomalies in the eastern and central equatorial Pacific (9° N–9° S, 90° W–180° W) where equatorial upwelling is most intense³¹. The temperature time series is derived from monthly analyses⁵⁰ smoothed twice with a 5-year running mean to filter out the seasonal cycle and year-to-year oscillations associated with ENSO. Anomalies are relative to 1950–99 averages.

Table 1 Water mass balance for the equatorial strip between 9° N and 9° S in the Pacific

	Year range			
	1956–65	1970–77	1980–89	1990–99
Surface layer divergence (Sv)	44.8 ± 5.3	51.7 ± 4.2	47.5 ± 4.5	40.7 ± 5.5
Interior ocean pycnocline convergence (Sv)	25.5 ± 3.0	27.0 ± 2.5	20.5 ± 1.6	14.0 ± 1.5
Western boundary convergence (Sv)	19.3 ± 6.1	24.7 ± 4.9	27.0 ± 4.8	26.7 ± 5.7
Equatorial upwelling (Sv)	40.9 ± 5.7	46.8 ± 4.6	42.1 ± 4.2	35.4 ± 4.8

Values shown are mean transport convergences and divergences, plus and minus one standard error. Convergence is computed as 9° S minus 9° N transports, and divergence as 9° N minus 9° S transports. Surface-layer divergence is based on Ekman transports reduced by opposing surface layer geostrophic transports.

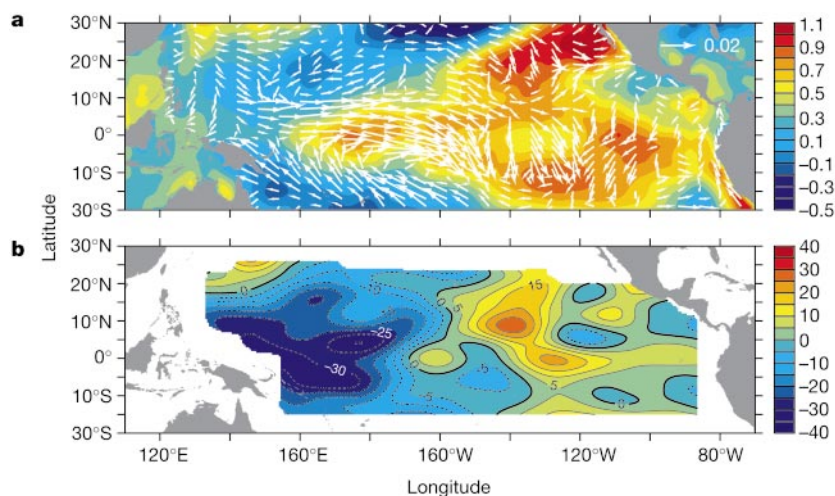


Figure 3 Decadal differences in the tropical Pacific between 1990–99 and 1970–77. **a**, Wind stress difference (in N m^{-2}) for 1990–99 minus 1970–77, based on the Florida State University wind product⁴⁴. Overplotted is the difference in sea surface temperature

(in $^{\circ}\text{C}$) the same time period⁵⁰. **b**, Depth difference (in m) of the 25.0 kg m^{-3} potential density surface for 1990–99 minus 1970–77.

occurred over the past several decades, they are likely to have been smaller than those for interior ocean geostrophic transports and surface-layer Ekman transports.

Comparison with Sverdrup theory

A complementary perspective on decadal variations in ocean pycnocline transports is available from Sverdrup theory²², the dynamics of which are fundamental to our understanding of large-scale oceanic circulation. This theory assumes that flow away from western boundaries is linear and in steady-state balance with surface wind stress forcing. The steady-state assumption is valid for the tropical oceans on decadal timescales, as the adjustment time in response to transient wind forcing at low latitudes is on the order of a few years²⁸.

Geostrophic transport per unit longitude based on Sverdrup theory is given by $M_g = (\rho_0 \beta)^{-1} [\nabla \times \tau + \beta \tau^x / f]$, where τ is wind stress vector and β is the change in the Coriolis parameter with latitude¹⁹. Using this expression, we computed meridional transport convergences, integrated from the eastern boundary to just outside the western boundary along 9°N and 9°S , for each wind stress product. Resultant ensemble averages and standard errors are $24.2 \pm 2.6 \text{ Sv}$ (1956–65), $27.1 \pm 6.8 \text{ Sv}$ (1970–77), $22.3 \pm 4.9 \text{ Sv}$ (1980–89) and $10.5 \pm 4.1 \text{ Sv}$ (1990–99). These estimates are comparable in magnitude to pycnocline transport convergences computed independently from hydrographic data, and exhibit a similar downward trend from the 1970s to the 1990s (Table 1).

Sverdrup theory provides no information about the vertical structures of geostrophic currents, so that exact comparison with observations would require integration of measured geostrophic flow over the entire water column. It is possible therefore that the good agreement between observed pycnocline transports and geostrophic transports predicted from Sverdrup theory may be fortuitous. However, the circulation in the tropical Pacific is most vigorous in the upper ocean²⁹, which leads us to conclude that the trend towards a weaker meridional overturning circulation from the 1970s to the 1990s is real.

Consequences for equatorial upwelling

Poleward flow in the surface Ekman layer extends to depths of about 50 m, whereas the pycnocline flows that we have identified extend to depths of hundreds of metres. The meridional overturning circulation is closed at low latitudes by upwelling, which brings cold pycnocline water to the surface where it once again makes contact

with the atmosphere. Most of the upwelling transport (90% or more) in the tropical Pacific occurs along the Equator in the eastern and central basin, with smaller amounts in the Costa Rica dome (centred near 10°N , 90°W) and in the coastal upwelling regime off Peru³⁰. Most of the upwelled water is of extratropical origin, although a portion is associated with shallow meridional recirculation cells confined to within a few degrees of the Equator^{25,29}. Some water upwelled near the Equator downwells around 3° – 4°N and 3° – 4°S , returning to the Equator at the base of the surface layer. These surface-layer recirculating transports are averaged out of our estimates, which are representative of the deeper water masses of subtropical origin flowing across 9°N and 9°S .

From mass continuity, we can infer that upwelling has transported pycnocline water toward the surface between 9°N and 9°S at an average rate of about 40 Sv over the past several decades. The magnitude of this transport corresponds well with previous estimates, although different latitude and longitude ranges for the various computations preclude exact comparison^{20,31,32}. Our results also indicate that these upwelling transports have varied decadal, decreasing by about 25% from 47 Sv in the 1970s to 35 Sv in the 1990s (Table 1). The reduction in upwelling is consistent with inferences from radiocarbon data of a decreased supply of pycnocline water to the surface layer in the eastern and central equatorial Pacific since the 1970s³³.

Implications for climate variability and climate change

The picture that emerges from this analysis is that the wind-driven meridional overturning circulation in the upper Pacific Ocean has been slowing down since the 1970s. This slowdown can account for the recent anomalous surface warming in the tropical Pacific, as the supply of cold pycnocline water originating at higher latitudes to feed equatorial upwelling has decreased. The Southern Hemisphere is responsible for about half of the observed decrease in equatorward pycnocline transport. Thus, perspectives on decadal variability limited to the Northern Hemisphere alone are incomplete. The fact that few studies have considered a role for the Southern Hemisphere ocean is presumably a consequence of limited data availability rather than a lack of decadal signal in the southern tropics and subtropics³⁴.

The oceanic and atmospheric processes that we have described work together so as to reinforce each other, similar to the positive feedbacks that occur during ENSO events. For example, weaker easterly trade winds in the equatorial Pacific would result in reduced

Ekman and geostrophic meridional transports, reduced equatorial upwelling, and warmer equatorial sea surface temperatures. Warmer surface temperatures in turn would alter patterns of deep atmospheric convection so as to favour weaker trade winds. If the system is to oscillate on decadal timescales, then delayed negative feedback mechanisms, one candidate for which involves planetary-scale ocean waves, must also be important^{4,16,35}.

Similarities in the spatial structures of the PDO and ENSO (both, for example, have phases that are characterized by warm tropics and a cool central North Pacific, and vice versa) have raised questions about the possible interaction between interannual- and decadal-timescale phenomena in the Pacific. In particular, since 1976–77 there have been fewer La Niña events, and more frequent, stronger, and longer-lasting El Niño events. Whether this recent change in the character of the ENSO cycle is a consequence or a cause of underlying decadal-timescale variability is unknown. It could be that the decadal changes in circulation described here operate independently from those that affect the ENSO cycle. If so, they would modify the background state on which ENSO develops, and thereby precondition interannual fluctuations to preferred modes of behaviour^{36,37}. Alternatively, the observed decadal changes may simply be the low-frequency residual of random or chaotic fluctuations in tropical ocean–atmosphere interactions that give rise to the ENSO cycle itself³⁸. In either case, a complete understanding of climate variability spanning interannual to decadal timescales in the Pacific basin will need to account for the slowly varying meridional overturning circulation between the tropics and subtropics.

Results of some numerical coupled ocean–atmosphere model studies suggest that the response of the tropical Pacific to greenhouse-gas forcing resembles a permanent El Niño-like condition^{35,39}. Thus, it is conceivable that climate fluctuations in the tropical Pacific over the past 50 years, including the recent slowdown of the meridional overturning circulation, may have been influenced by global warming as well as by natural variability. Unfortunately, separating out the putative effects of anthropogenically forced climate change from natural variations is not possible with relatively short data records. Our analysis can, however, provide an important dynamical constraint for model studies that attempt to simulate recent observed decadal changes in the Pacific basin.

Finally, we note that the equatorial Pacific is the largest oceanic source of CO₂ to the atmosphere^{40,41}. It also accounts for roughly 20% of the photosynthetically mediated new biological production in the world's oceans³⁰. Upwelling of nutrient- and carbon-rich water to the sea surface is the principal reason why the equatorial Pacific is so important in global biogeochemical cycles. Consistent with a decrease in upwelling in the 1990s relative to earlier decades, the outgassing of CO₂ from the equatorial Pacific Ocean has decreased from the 1980s to the 1990s^{40,41}, potentially reducing the rate at which CO₂ would otherwise have accumulated in the atmosphere⁴¹. It is also likely (though difficult to confirm for lack of adequate data) that the supply of nutrients to the surface layer of the equatorial Pacific has decreased from the 1970s to the 1990s. A reduced nutrient supply would favour a reduction in biological productivity in the equatorial Pacific, which in turn would affect overall global rates of carbon fixation. □

Methods

Processing of hydrographic data

Hydrographic data were obtained from the World Ocean Data Base⁴² augmented with data from archives of the Pacific Marine Environmental Laboratory¹⁷. A total of 114,691 casts between 20° S and 50° N reach 900 dbar (1 dbar = 0.992 m depth), which is the reference level of our geostrophic velocity estimates. The number of casts for defining water mass properties on shallow density surfaces is greater (for example, 215, 993 casts on the 25.0 kg m⁻³ isopycnal shown in Fig. 1). By decade, the number of casts to 900 dbar is 11,584 (1950–59), 21,779 (1960–69), 24,920 (1970–79), 36,110 (1980–89) and 20,298 (1990–99).

Data coverage is relatively sparse in the Southern Hemisphere, which limits our analysis

to latitudes north of 20° S. Data are also unevenly distributed spatially and temporally, potentially aliasing interannual, seasonal and shorter timescale and space-scale fluctuations into the decadal climate signals in which we are interested. Despite these sampling limitations, the data are adequate to define robust features of the mean circulation and its large-scale variability averaged over periods of 7–10 years.

Data were binned in boxes of 0.5° latitude by 0.5° longitude, then gridded objectively by decade on density surfaces. We assumed a gaussian correlation structure function with a half-width of 5° latitude by 15° longitude in the interior ocean, and 5° latitude by 5° longitude near the western boundary. The distribution of hydrocast data on the larger of these scales is shown in Fig. 1b. Our choice of 900 dbar as a reference level for geostrophic velocity computations is consistent with previous analyses in the tropical and subtropical Pacific¹⁷. This level represents a compromise between optimizing data coverage (which is better near to the surface) and capturing weak horizontal flows below 900 dbar. Choosing alternative reference depths (for example, 600 dbar or 1,200 dbar) leads to only minor differences in pycnocline transport values (0.5–1 Sv).

The zonal integral of interior ocean meridional geostrophic transport is effectively the difference between depth-integrated dynamic height at opposite ends of the basin. To estimate the random sampling error for these transports, we performed Monte Carlo simulations on subsets of data near the eastern and western limits of the basin for each decadal period. Typically hundreds of data pairs were available for these assessments within several degrees of latitude and longitude around the end points of the transport calculation. Results are shown as one standard error in text, figures and tables.

Ekman and Sverdrup transports

We use four wind products to compute Ekman and Sverdrup transports: the Comprehensive Ocean Atmosphere Data Set (COADS) winds for 1950–93⁴³, the Florida State University (FSU) winds for 1961–99⁴⁴, the National Centers for Environmental Prediction (NCEP) winds for 1958–99⁴⁵, and the European Centre for Medium Range Weather Forecasts (ECMWF)⁴⁶ winds for 1979–99. COADS provides a stress product directly; FSU, NCEP and ECMWF wind analyses were converted to wind stress using essentially the same algorithm as used in COADS. We required a minimum of 5 years in any given period to compute a decadal average.

There are shortcomings in these (and other) wind products in terms of potential systematic biases related to data coverage, analysis procedures, or both^{43,47}. Therefore, rather than rely on any one product alone, we formed ensemble means of the different products to draw out significant climate signals common to all. This procedure also provides a measure of the uncertainty (in the form of statistically derived standard errors) in our representation of those signals.

Data shown in Table 1

Interior ocean pycnocline transports relative to 900 dbar were derived from hydrographic data, and convergence was computed as the difference of 9° S minus 9° N values. Ekman transports were adjusted downward for measured equatorward geostrophic flows in the surface layer of the interior ocean, and divergence computed as the difference of 9° N minus 9° S. Western boundary current transport convergences represent the difference between surface-layer divergences and interior ocean pycnocline convergences, assuming that mass is conserved in the upper ocean. Upwelling transports between 9° N and 9° S were computed by assigning 80% of the western boundary current transport convergence to the pycnocline layer and 20% to the surface layer, based on observed mean vertical structures of these currents.

Western boundary current transport convergences in Table 1 represent lower bounds, because these currents also supply upper pycnocline water to the Indonesian throughflow²⁷. However, those additional sverdrups of boundary current transport required to feed the throughflow, since they simply leave the basin, do not affect our equatorial upwelling computations. Moreover, we estimate that on decadal timescales the Indonesian throughflow has remained constant to within about ±1 Sv on the basis of calculations using Godfrey's island rule⁴⁸ forced by global NCEP⁴⁵ and ECMWF⁴⁶ wind fields. Coupled ocean–atmosphere model integrations indicate a similar range of possible decadal-timescale variations in Indonesian throughflow⁴⁹. These fluctuations are much smaller than those associated with the meridional overturning circulation in the upper Pacific.

Received 7 August; accepted 7 December 2001.

1. Minobe, S. Spatio-temporal structure of pentadecadal variability over the North Pacific. *Prog. Oceanogr.* **47**, 381–408 (2000).
2. Trenberth, K. E. & Hurrell, J. W. Decadal atmosphere-ocean variations in the Pacific. *Clim. Dyn.* **9**, 303–319 (1994).
3. Mantua, N. J., Hare, S. J., Zhang, Y., Wallace, J. M. & Francis, R. C. A Pacific interdecadal oscillation with impacts on salmon production. *Bull. Am. Meteorol. Soc.* **76**, 1069–1079 (1997).
4. Zhang, Y., Wallace, J. M. & Battisti, D. S. ENSO-like interdecadal variability: 1900–93. *J. Clim.* **10**, 1004–1020 (1997).
5. Hare, S. R. & Mantua, N. Empirical evidence for North Pacific regime shifts in 1977 and 1989. *Prog. Oceanogr.* **47**, 103–145 (2000).
6. Cayan, D. R., Kammerdiener, S. A., Dettinger, M. D., Caprio, J. M. & Peterson, D. H. Changes in the onset of spring in the western United States. *Bull. Am. Meteorol. Soc.* **82**, 399–415 (2001).
7. Miller, A. J. & Schneider, N. Interdecadal climate regime dynamics in the North Pacific Ocean: theories, observations, and ecosystem impacts. *Prog. Oceanogr.* **47**, 355–379 (2000).
8. Latif, M. & Barnett, T. P. Cause of decadal variability over the North Pacific and North America: Dynamics and predictability. *J. Clim.* **9**, 2407–2423 (1996).
9. Lindsey, B. K., Wellington, G. M. & Schrag, D. P. Decadal sea surface temperature variability in the subtropical South Pacific from 1726 to 1997 A. D. *Science* **290**, 1145–1148 (2001).

10. Garreaud, R. E. & Battisti, D. S. Interannual (ENSO) and interdecadal (ENSO-like) variability in the Southern Hemisphere tropospheric circulation. *J. Clim.* **12**, 2113–2123 (1999).
11. Gu, D. & Philander, S. G. H. Interdecadal climate fluctuations that depend on exchanges between the tropics and extratropics. *Science* **275**, 805–807 (1997).
12. Schneider N., Miller, A. J., Alexander, A. & Deser, C. Subduction of decadal North Pacific temperature anomalies: Observations and dynamics. *J. Phys. Oceanogr.* **29**, 1056–1070 (1999).
13. Schneider, N. *et al.* Pacific thermocline bridge revisited. *Geophys. Res. Lett.* **26**, 1329–1332 (1999).
14. Liu, Z. & Shin, S.-I. On thermal ventilation of active and passive tracers. *Geophys. Res. Lett.* **26**, 357–360 (1999).
15. Nonaka, M., Xie, S.-P. & Takeuchi, K. Equatorward spreading of a passive tracer with application to North Pacific interdecadal temperature variations. *J. Oceanogr.* **56**, 173–183 (2000).
16. Kleeman, R., McCreary, J. P. & Klinger, B. A mechanism for generating ENSO decadal variability. *Geophys. Res. Lett.* **26**, 1743–1746 (1999).
17. Johnson, G. C. & McPhaden, M. J. Interior pycnocline flow from the subtropical to the equatorial Pacific Ocean. *J. Phys. Oceanogr.* **29**, 3073–3089 (1999).
18. Huang, B. & Liu, Z. Pacific subtropical-tropical thermocline water exchange in the National Centers for Environmental Prediction ocean model. *J. Geophys. Res.* **104**, 11065–11076 (1999).
19. McPhaden, M. J. & Fine, R. A. A dynamical interpretation of the tritium maximum in the central equatorial Pacific. *J. Phys. Oceanogr.* **18**, 1454–1457 (1988).
20. Quay, P. D., Stuiver, M. & Broecker, W. S. Upwelling rates for the equatorial Pacific Ocean derived from the bomb ^{14}C distribution. *J. Mar. Res.* **41**, 769–792 (1983).
21. Fine, R. A., Maillet, K. A., Sullivan, K. F. & Willey, D. Circulation and ventilation flux of the Pacific Ocean. *J. Geophys. Res.* **106**, 22159–22178 (2001).
22. Sverdrup, H. U. Wind-driven currents in a baroclinic ocean: with application to the equatorial currents of the eastern Pacific. *Proc. Natl Acad. Sci. USA* **33**, 318–326 (1947).
23. Liu, Z. & Huang, B. Cause of tropical Pacific warming trend. *Geophys. Res. Lett.* **27**, 1935–1938 (2000).
24. Wallace, J. M. *et al.* On the structure and evolution of ENSO-related climate variability in the tropical Pacific: Lessons from TOGA. *J. Geophys. Res.* **103**, 14241–14259 (1998).
25. McCreary, J. P. & Lu, P. On the interaction between the subtropical and the equatorial oceans: The subtropical cell. *Phys. Oceanogr.* **24**, 466–497 (1994).
26. Ralph, E. A. & Niiler, P. P. Wind-driven currents in the tropical Pacific. *J. Phys. Oceanogr.* **29**, 2121–2129 (1999).
27. Gordon, A. L., Susanto, R. D. & Field, A. Throughflow within the Makassar Strait. *Geophys. Res. Lett.* **26**, 3325–3328 (1999).
28. Philander, S. G. H. Variability of the tropical oceans. *Dyn. Atmos. Oceans* **3**, 191–208 (1979).
29. Johnson G. C., McPhaden, M. J. & Firing, E. Equatorial Pacific Ocean horizontal velocity, divergence, and upwelling. *J. Phys. Oceanogr.* **31**, 839–849 (2001).
30. Chavez, F. P. & Toggweiler, J. R. in *Upwelling in the Ocean: Modeling Processes and Ancient Records* (eds Summerhayes, C. P. *et al.*) 313–320 (Wiley, New York, 1995).
31. Wyrtki, K. An estimate of equatorial upwelling in the Pacific. *J. Phys. Oceanogr.* **11**, 1205–1214 (1981).
32. Meinen, C. S., McPhaden, M. J. & Johnson, G. C. Vertical velocities and transports in the equatorial Pacific during 1993–1999. *J. Phys. Oceanogr.* **31**, 3230–3248 (2001).
33. Guilderson, T. P. & Schrag, D. P. Abrupt shift in subsurface temperatures in the tropical Pacific associated with changes in El Niño. *Science* **281**, 240–243 (1998).
34. Chang, P., Giese, B. S., Ji, L., Seidel, H. F. & Wang, F. Decadal change in the south tropical Pacific in a global assimilation analysis. *Geophys. Res. Lett.* **28**, 3461–3464 (2001).
35. Knutsen, T. R. & Manabe, S. Model assessment of decadal variability and trends in the tropical Pacific Ocean. *J. Clim.* **11**, 2273–2296 (1998).
36. Federov, A. V. & Philander, S. G. H. Is El Niño changing? *Science* **288**, 1997–2001 (2000).
37. Urban, F. E., Cole, J. E. & Overpeck, J. T. Influence of mean climate change on climate variability from a 155-year tropical Pacific coral record. *Nature* **407**, 898–993 (2000).
38. Kirtman, B. P. & Schopf, P. S. Decadal variability in ENSO predictability and prediction. *J. Clim.* **11**, 2843–2822 (1998).
39. Meehl, G. A. & Washington, W. M. El Niño-like climate change in a model with increased CO_2 concentrations. *Nature* **382**, 56–60 (1995).
40. Feely, R. A., Wanninkhof, R., Takahashi, T. & Tans, P. Influence of El Niño on the equatorial Pacific contribution to atmospheric CO_2 accumulation. *Nature* **398**, 597–601 (1999).
41. Feely, R. A. *et al.* Seasonal and interannual variability of CO_2 in the equatorial Pacific. *Deep-Sea Res.* (in the press).
42. Conkright, M. E. *et al.* *World Ocean Database 1998 Version 2.0* (National Oceanographic Data Center Internal Report 14, National Oceanic and Atmospheric Administration, Silver Spring, Maryland, 1999).
43. da Silva, A. M., Young, C. C. & Levitus, S. *Atlas of Surface Marine Data Vol. 1, Algorithms and Procedures* (NOAA Atlas NESDIS 6, US Dept. of Commerce, Washington DC, 1994).
44. Goldenberg, S. B. & O'Brien, J. J. Time and space variability of tropical Pacific wind stress. *Mon. Weath. Rev.* **109**, 1190–1207 (1981).
45. Kalnay, E. *et al.* The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**, 437–471 (1996).
46. European Centre for Medium-Range Weather Forecasts – ECMWF Data Archive (<http://www.ecmwf.int/services/data/archive/>) (February 2001).
47. Hackert, E. C., Busalacchi, A. J. & Murtugudde, R. A. A wind comparison study using an ocean general circulation model for the 1997–98 El Niño. *J. Geophys. Res.* **106**, 2345–2362 (2001).
48. Godfrey, J. S. A Sverdrup model of the depth-integrated flow for the World Ocean allowing for island circulations. *Geophys. Astrophys. Fluid Dyn.* **45**, 89–112 (1989).
49. Schneider, N. The Indonesian throughflow and the global climate system. *J. Clim.* **11**, 676–689 (1998).
50. Smith, T. M., Reynolds, R. W. & Ropelewski, C. F. Optimal averaging of seasonal sea surface temperatures and associated confidence intervals. *J. Clim.* **7**, 949–964 (1994).

Acknowledgements

We thank J. Potemra for advice on Indonesian throughflow calculations. We also thank W. Kessler, G. Johnson and R. Kleeman for comments on an earlier version of this manuscript. This work was supported by the NOAA Office of Oceanic and Atmospheric Research and by the Joint Study for the Atmosphere and Ocean (JISAO) at the University of Washington.

Correspondence and requests for materials should be addressed to M.J.M. (e-mail: mcphaden@pmel.noaa.gov).

Visual predators select for crypticity and polymorphism in virtual prey

Alan B. Bond* & Alan C. Kamil*†

*Nebraska Behavioral Biology Group, School of Biological Sciences; and †Department of Psychology, University of Nebraska, Lincoln, Nebraska 68588-0118, USA

Cryptically coloured animals commonly occur in several distinct pattern variants. Such phenotypic diversity may be promoted by frequency-dependent predation, in which more abundant variants are attacked disproportionately often, but the hypothesis has never been explicitly tested. Here we report the first controlled experiment on the effects of visual predators on prey crypticity and phenotypic variance, in which blue jays (*Cyanocitta cristata*) searched for digital moths on computer monitors. Moth phenotypes evolved via a genetic algorithm in which individuals detected by the jays were much less likely to reproduce. Jays often failed to detect atypical cryptic moths, confirming frequency-dependent selection and suggesting the use of searching images, which enhance the detection of common prey. Over successive generations, the moths evolved to become significantly harder to detect, and they showed significantly greater phenotypic variance than non-selected or frequency-independent selected controls.

Polymorphic cryptic coloration has been recorded in a variety of prey species, including grasshoppers¹, homoptera², mantids³ and bivalves⁴, but it is particularly evident in the noctuid moths that rest concealed on tree trunks and other vegetation during the day^{5,6}. Roughly 45% of North American moths of the genus *Catocala* are polymorphic, with some species occurring in as many as nine different forms⁷. In other noctuid genera, the adults are “massively polymorphic”, varying continuously in appearance over an exceedingly broad range^{8–10}.

Frequency-dependent selection, in which more common prey types are attacked disproportionately often, has been assumed to play a role in maintaining colour pattern polymorphism^{11–12}. If predators respond to changes in relative abundance by switching away from once common prey when they become rare, such ‘apostatic selection’ could prevent rare prey types from being eliminated and thereby maintain balanced numbers of the different morphs. Although many ecological mechanisms can produce frequency-dependent selection, polymorphism in cryptic prey species probably reflects the fact, first noted by Poulton¹³ in 1890, that it is difficult and time-consuming to hunt for several different items simultaneously. Consequently, visual predators tend to search at any moment for only a limited number of prey types, focusing on the distinctive features of common ones and effectively overlooking the others, a phenomenon known as “hunting by searching image”¹⁴. Recent studies have confirmed the occurrence of searching image effects, demonstrating that the ability to detect cryptic prey increases when items of similar appearance are encountered in rapid succession and that a search for one cryptic prey type can actively interfere with the detection of alternative forms^{15–18}.

Because the dynamic interplay between predator behaviour and prey appearance is difficult to evoke under controlled conditions, much of the earlier empirical support for frequency-dependent selection of cryptic prey was fragmentary and indirect¹⁹. To circumvent this constraint, we developed a ‘virtual ecology’, in which captive blue jays hunt for artificial, digital moths on computer displays. The preparation allows realistic, repeatable experimental investigations of the ecological processes involved in prey evolution. Our previous work with this system has shown that predators searching for a set of fixed prey types do exhibit frequency-dependent selection, that such selection does serve to maintain stable prey polymorphism, and that the effect is almost certainly attributable to hunting by searching image^{18,19}.

Because frequency-dependent selection operates against common, familiar phenotypes, it should also lead to the proliferation of new,

disparate phenotypes in an initially monomorphic prey population^{4,20–22}. This prediction has been subjected to only limited experimental testing. In our previous study of virtual ecology, frequency-dependent selection produced stable dynamics among a set of three morphs, but the consequences of adding new morphs seemed to depend on how difficult they were to detect¹⁸. In one case, the new morph was similar in crypticity to those in the initial population, and a stable configuration of four prey types was produced. A second new morph, however, was so cryptic and distinctive that few of them were ever discovered. This new morph came to dominate the population completely and drove all of the others to extinction. The dynamics that one might expect from natural genetic changes in prey populations are therefore unclear. By extending our virtual ecology paradigm through the addition of a genetic algorithm, we have found a means of addressing this issue and testing whether frequency-dependent selection not only maintains but actively encourages the evolution of phenotypic diversity.

Virtual ecology

Our methods derive from a naturally occurring predator–prey system. In eastern North America, blue jays commonly prey on cryptically coloured noctuid moths during the daytime, discovering the insects even while they are resting motionless on tree trunks⁵. Jays exhibit similarly impressive detection abilities when required to locate moths in projected images in the laboratory, providing a convincing emulation of natural foraging behaviour^{23–25}. To furnish our jays with a functional virtual ecology, we generated synthetic, digital moths—bilaterally symmetrical triangles about 6 mm high—and displayed them to the birds overlaid on a complex, granular background (Fig. 1)^{18,19}. To avoid potential problems with avian colour perception, the displays were constructed using a 64-level grey scale²⁶.

Each moth phenotype was constructed from specifications in a virtual chromosome, through an algorithm derived from salient features of the genetics of lepidopteran wing patterns^{27–30}. These included genes that coded for individual patches of pattern elements, genes that produced global changes in brightness or contrast, and linkage mechanisms that protected favourable genotypes from being broken up by recombination. As in real moths, phenotypic traits were polygenic, in that the intensity of any given pixel was the result of additive interactions among a large number of loci.

Our experimental design was similar to that of more traditional selection studies, contrasting the selective effects of jay predation in

several experimental lineages with the results of control treatments in which selection was eliminated or constrained. In each of our experimental and control lineages, the moth population was maintained at a constant level of 200 individuals over 100 successive, non-overlapping generations. Reproduction entailed choosing two individual chromosomes according to the selection algorithm and recombining them into a single offspring genome. The offspring

were then subjected to a mutation process that randomly inverted individual bits. To produce a new generation, these selection, recombination and mutation steps were repeated until a full array of 200 progeny had been obtained. The previous population was then discarded and replaced with the new individuals.

An initial, monomorphic parental population with moderate phenotypic variability (Fig. 1a) was generated from a template provided by a moth image from an earlier experiment¹⁹. This population was then hunted by four jays, each of which was required to search for the moths presented one at a time on a computer display^{18,19}. Each moth in the population was presented to one of the jays once in the course of a daily session. Half of these daily trials, on a random basis, contained a moth embedded somewhere in the display, while the other half showed only the background. If a jay correctly detected a moth and pecked at it, the bird was rewarded with a food pellet; if it did not find a moth, the bird pecked a central green circle to initiate the next trial.

At the end of each day, the accuracy and latency of the birds' responses were scored and entered into the selection algorithm, which favoured those moths that were most difficult for the jays to detect. The moth population was allowed to reproduce, and the resulting generation was presented to the jays on the succeeding day. Beginning each time with the same parental population, we produced three successive experimental lineages, continuing selection in each case out to the F_{100} generation. Between runs, the jays were given 30 days of exposure to the parental population under stationary, non-evolving conditions to return them to a consistent baseline.

To test whether observed changes in crypticity and phenotypic variance in the experimental lines were statistically meaningful, the results were contrasted with those from two sets of control lineages, one employing no selective agency and the other using frequency-independent directional selection based on parameters derived from global aspects of the jays' behaviour. In both of the control treatments, we used the same population size, the same initial parental population, the same backgrounds and the same mutation rate as in the experimental trials. In the non-selected lineages, however, the moths were never presented to jays, and the probability of being chosen to breed was uniform across the moth population, irrespective of phenotype. This provided a control for the occurrence of directional selection for crypticity in the experimental treatments.

The second control was designed to assess our primary hypothesis, that frequency-dependent selection promotes increased phenotypic diversity. This required comparison to lineages in which selection was independent of the frequency of particular phenotypes, but was otherwise similar in intensity and direction to that produced in the experimental lines. For these control lineages, therefore, we determined the functional relationship between detection and crypticity for the jays, averaging over all of the results in the experimental lines. This function was then used to determine the probability of a moth's being chosen to breed. In the experimental trials, the jays appeared to allocate their searching effort at least in part on the basis of the average degree of difficulty they experienced in finding the moths: Cryptic moths were detected more accurately and rapidly when the rest of the individuals in the population were cryptic than when the rest of them were relatively conspicuous ($r^2 = 0.9$, student's $t(1) = 6.74$, $P < 0.002$).

This 'search rate'³¹ or 'caution'¹⁵ effect has been observed in other experimental studies of foraging behaviour, and although caution cannot produce frequency-dependent predation when prey are presented in a randomized sequence³², it provided a solid basis for an alternative, frequency-independent control. The selection algorithm in the frequency-independent lineages was based on the empirical relationships between crypticity and detection performance in the experimental trials and fitted to the observed effects of mean population crypticity. We generated 200 selection lines of

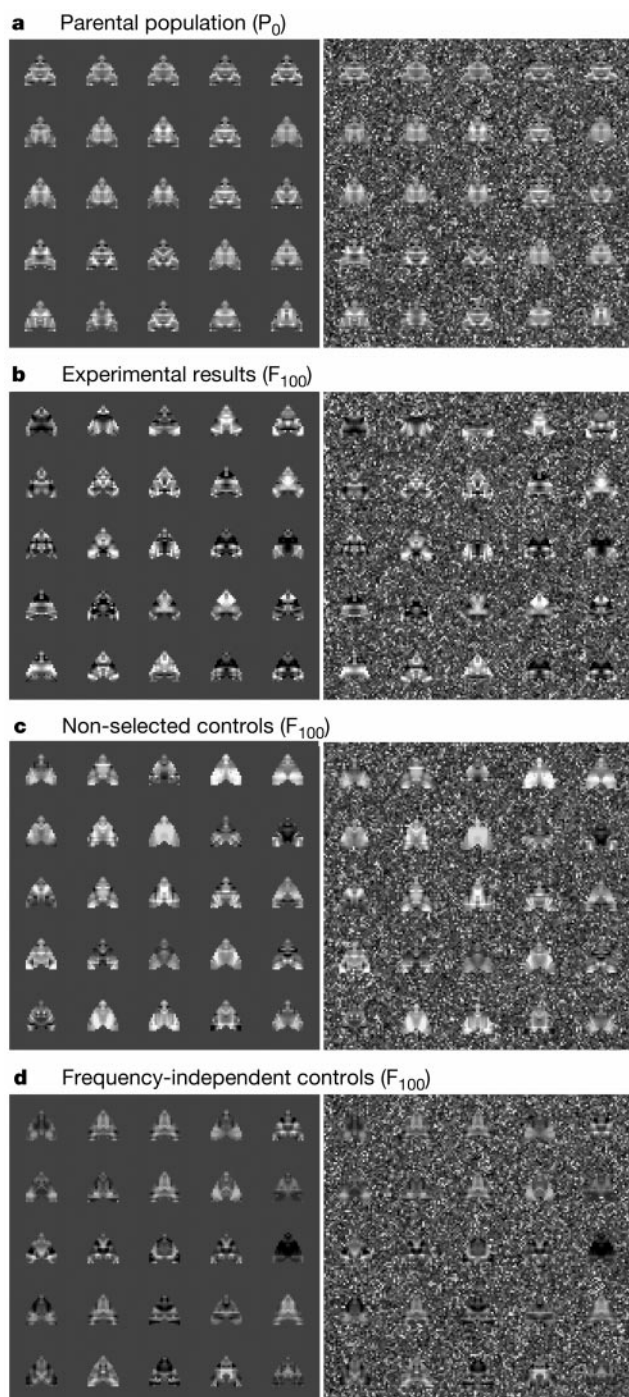


Figure 1 Samples of digital moths, shown on uniform grey (left) and cryptically textured (right) backgrounds, from the parental population, P_0 (a), and from the F_{100} generations from the non-selected lines (b), the frequency-independent selection lines (c), and the experimental lines selected by the jays (d). We note that moths from the experimental lines are generally more cryptic than those in the non-selected lines and more variable in appearance than those in the lines subjected to frequency-independent selection.

100 successive generations using the non-selected and frequency-independent control treatments to provide controls for the selective effects of jay predation.

Frequency-dependent selection

In our previous studies of virtual ecology, which used morphs of fixed appearance, we found clear evidence of frequency-dependent selection and were able to show that such selection served to maintain a stable prey polymorphism. To confirm that a similar process was also operating on the continuously variable morphs in our current experiment, we examined the factors influencing detection accuracy in the experimental lines. Subjects varied in their ability to detect cryptic prey, with two of the birds consistently showing higher accuracy than the others ($F_{3,400} \geq 49.67$, $P < 0.0001$). Mean accuracy across subjects remained almost constant over the course of each experimental lineage ($r^2 = 0.04$; $t(1) = 1.18$, NS) but increased significantly across successive lineages ($F_{2,30} = 10.48$, $P < 0.001$). The pattern of variation of accuracy over time was quite complex, exhibiting a strong, regular oscillation with a fundamental period of about 10 generations (minimum Durbin–Watson statistic, $DW(10) = 1.503$, $P < 0.001$). Oscillations in accuracy of roughly the same periodicity were also observed in our previous work on selection with fixed morphs¹⁹, where they were linked to the rise and fall in abundance of the corresponding prey types. Although we cannot readily categorize phenotype abundance here, the accuracy oscillation presumably derives from a similar unstable feedback between the predators' focal search for a limited number of visual features and the resulting selective pressure against prey that carry those features.

More cogent evidence of the selective mechanisms can be obtained from close examination of the effects of trial sequence. If the jays were using searching images in their pursuit of digital moths, cases in which several moths of similar appearance were presented in succession would be expected to produce a temporary enhancement in detection, while moths of disparate appearance should be more likely to be overlooked¹⁸. Stepping through the results, trial by trial, we extracted the phenotypic distance between each moth and the most recent previously detected moth. We corrected this dissimilarity measure for the mean dissimilarity and crypticity among all moths in each population, producing a relative measure of successional dissimilarity for all individuals in all three experimental lines.

For analysis, we pooled the data from the three lines and divided them into three groupings on the basis of moth crypticity. Within each grouping, trials were sorted by relative dissimilarity and blocked into sets of 100. Detection accuracy in each set was then analysed as a function of dissimilarity and crypticity grouping (Fig. 2). There was a strong, significant effect of crypticity level on accuracy ($F_{2,554} = 709$, $P < 0.0001$): the mean accuracy on trials with low-crypticity moths was 0.92, with medium-crypticity moths 0.89, and with high-crypticity moths 0.76. The relationship between accuracy and dissimilarity was noticeably affected by crypticity. At low- and medium-crypticity levels, the slopes of the regression of accuracy on dissimilarity were not significant ($r^2 \leq 0.021$; $t(1) \leq 1.97$, NS). The high crypticity grouping, on the other hand, showed a strong negative relationship between accuracy and dissimilarity ($r^2 = 0.35$; $t(1) = -10.3$, $P < 0.0001$).

That atypical cryptic prey items should be more difficult to detect is precisely what one would expect from the operation of searching images, and it substantiates our assumption that moths in the experimental lineages were subjected to frequency-dependent selection. Sequential dissimilarity had no evident effect on the detection of relatively conspicuous prey, however, replicating numerous previous findings that searching image mainly plays a role in the detection of cryptic stimuli^{15–17,33}. In addition, to our knowledge these results provide the first evidence of searching image effects where prey appearance was continuously variable;

all previous studies have employed a limited number of fixed, easily distinguishable prey types.

Selection effects on crypticity

Our next concern was whether jay predation resulted in directional selection for crypticity over the course of the study, that is, whether the moths evolved to become more difficult to detect. Mean crypticity indices from the three experimental lines rose from 0.47 in the parental generation to an average of 0.60 in the last five generations, an increase of 29% (Fig. 3). Crypticity also increased in the non-selected control lines, apparently as a result of initial conditions: the parental population was slightly less cryptic than the long-term average of non-selected lines, causing a gradual drift towards greater crypticity in the first 100 generations. All three experimental lines, however, did display a significantly higher level of crypticity in the F_{100} generation than did the non-selective controls (Fig. 3a, $F_{1,40596} \geq 11.51$, $P < 0.0001$).

Jay predation thus resulted in unequivocal, directional selection for increased crypticity in the digital moth populations, as can readily be confirmed in even small, arbitrary samples of individuals

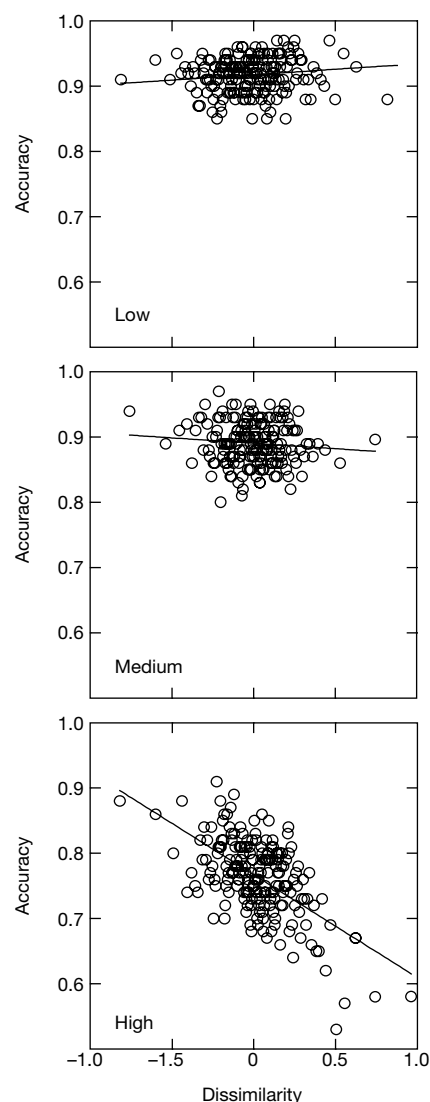


Figure 2 Detection accuracy of blocks of 100 trials as a function of crypticity of the target moth and dissimilarity between the target and the last previous correctly detected moth. Dissimilarity has been adjusted to remove the influence of mean phenotypic variance and mean population crypticity. Individual moth crypticity, c , increases from the top to the bottom panel: low ($c \leq 0.45$), medium ($0.45 < c \leq 0.6$), and high ($c > 0.6$). Regression lines indicate the relationship between accuracy and dissimilarity within crypticity levels.

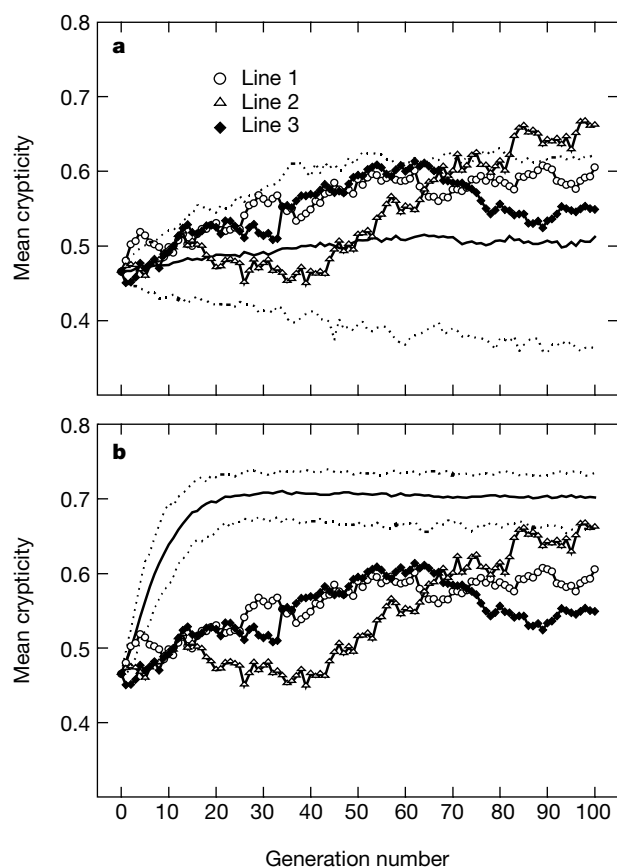


Figure 3 Changes in mean crypticity across successive generations in the three experimental lines (plotted with symbols), contrasted with the distribution of values from the two sets of control lines. Non-selected lines form the control group in **a**; the control in **b** was produced by frequency-independent selection for crypticity based on parameters derived from global aspects of the jays' behaviour. Graphs display medians (solid lines) and 95% confidence limits (dotted lines) from 200 replicate control lines. Crypticity increased across generations to some degree in all three treatments; the increase was greatest for the frequency-independent controls and least for the non-selected lines.

from the hundredth generation (Fig. 1b, c). Selection in favour of individuals that resemble the background has been invoked as the probable cause of cryptic coloration in prey species for over a century^{13,34}, and there have been numerous demonstrations that predators preferentially feed on more conspicuous prey items^{3,35–37}. Our study is, however, the only work other than Endler's research on colour-pattern selection in guppies³⁸ that has shown significant directional selection by predators over multiple successive prey generations when compared with a non-selected control.

Mean crypticity increased more gradually in the experimental lines than in the frequency-independent controls and was still significantly lower in the F_{100} generation (Figs 1b, d and 3b; $F_{1,40596} \geq 28.7$, $P < 0.0001$). This should not be surprising, as frequency-dependent selection should enable infrequent, moderately cryptic phenotypes to persist long after a directional, frequency-independent process would have led to their extinction. In addition, if the jays were focusing their search on specific, salient features of individual moth phenotypes, they may often have overlooked even relatively conspicuous individuals, accentuating the disparity between experimental and control treatments.

Selection effects on phenotypic variance

In our final analysis, the role of frequency-dependent selection in promoting phenotypic variance was tested by comparing the diversity of phenotypes in the experimental lines with that in the control lineages. Phenotypic diversity increased in the three experimental

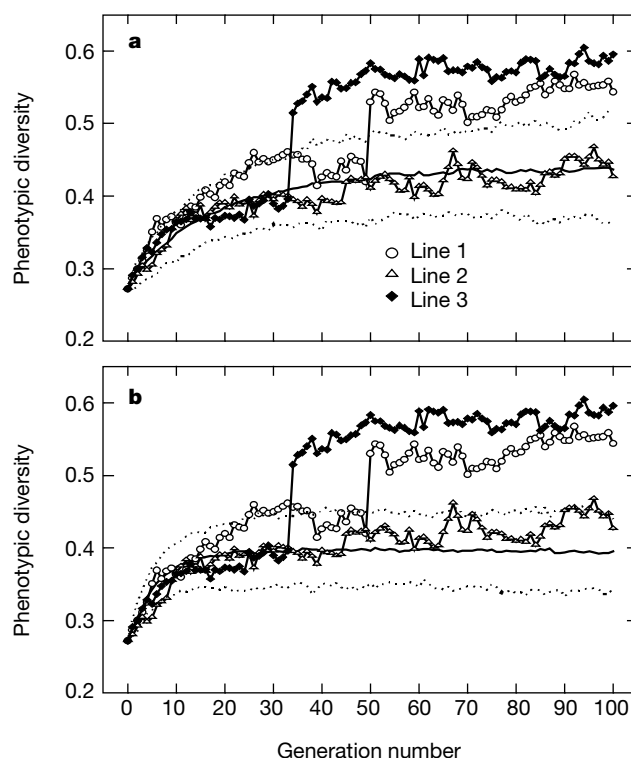


Figure 4 Changes in phenotypic variance of the digital moth population in the three experimental lines (plotted with symbols), contrasted with the distribution of values from the two sets of control lines. Non-selected lines form the control group in **a**; the control in **b** was produced by frequency-independent selection for crypticity based on parameters derived from global aspects of the jays' behaviour. Graphs display medians (solid lines) and 95% confidence limits (dotted lines) from 200 replicate control lines. Phenotypic variance increased to some degree in all three treatments, but the increase was greater in the experimental lines than in the controls. Experimental lines 1 and 3 each exhibited an abrupt shift to a higher level of phenotypic variance at some point in the course of selection trials. In each case, the shift appeared to have been produced by the explosive spread of mutant regulatory genes affecting global levels of brightness or contrast.

lines from 0.27 in the parental population to an average of 0.53 in the last five generations, an increase of 95% (Fig. 4). Although the diversity of phenotypes also increased in both control treatments, all three experimental lines showed significantly higher diversity in the F_{100} generation than did the frequency-independent controls (Figs 1b, d and 4b; $F_{1,40596} \geq 21.44$, $P < 0.0001$), and two of the three lines were even higher than the non-selected controls (lines 1 and 3, Figs 1b, c and 4a; $F_{1,40596} \geq 57.67$, $P < 0.0001$). The significantly higher variance exhibited in the experimental lines thus provides the first direct evidence of the polymorphic effects of frequency-dependent selection.

The observed increase in phenotypic diversity does not meet Ford's restrictive definition of polymorphism³⁹, in that the morphs were neither discontinuously variable nor controlled by a single switch gene. It would be surprising, however, if we had obtained discontinuous phenotypic variation in the absence of significant background heterogeneity³⁵. Instances of similarly phenotypically diverse, or "massively polymorphic", cryptic species are by no means uncommon in the literature, having been recorded in brittle-stars⁴⁰, bivalves⁴, grasshoppers⁴¹, fish^{35,38} and noctuid moths^{8–10}. One feature that these disparate organisms have in common is high levels of local abundance, which ensures that predators will experience short time delays between successive encounters with the species. Short interprey intervals may tend to accentuate the influence of searching images, relative to other selective factors that affect prey colour patterns⁴².

High phenotypic diversity can therefore be selected for in abundant prey species living in fine-grained habitats. In the next stage of our explorations of virtual ecology, we will introduce a range of heterogeneous backgrounds to examine the evolution of more classical, discontinuous polymorphisms. This will allow a direct test of the hypothesis that high phenotypic diversity is, as has been suggested⁴³, a step along the evolutionary path toward more discrete and genetically integrated polymorphisms. □

Methods

Subjects

Four blue jays were captured in the field as nestlings and hand-reared in the laboratory. They were initially trained to peck at a single, invariant moth on a uniform grey background. Subsequently, negative trials without moths were added to the daily sessions, and the jays were given extensive discrimination training on a set of six fixed moth phenotypes. When the birds reached a criterion of 70–80% correct responses, we began to embed the moths in backgrounds of gradually increasing levels of difficulty. Once all four subjects could find the fixed set of moths at criterion levels on the experimental backgrounds, we introduced the birds to the parental population under non-evolving conditions. Stable performance by all subjects of at least 70% correct was required to begin selection trials.

Genetic algorithms

Moth phenotypes were developed from specifications in a virtual haploid chromosome, a string of 117 bytes. The wing pattern was encoded by 18 loci, each consisting of five bytes that defined an elliptical wing patch by specifying its location, its orientation, the ratio of its axis lengths, the spread of intensity from its centre, and its base intensity. The value of each pixel in the phenotype was determined by the additive result of multiple overlapping patches, which provided some redundancy and assisted in maintaining the level of genetic variance in the population in spite of repeated generations of selection.

The chromosome was divided into nine linkage groups, each consisting of two patch loci and a regulatory locus. Each regulatory locus included one gene for brightness, one for contrast, and one for the probability of crossing over. Once the primary pattern was decoded from the patch loci, the developmental algorithm calculated the mean values of the brightness and contrast genes and modified the final image accordingly. This feature enabled fine adjustments in the colour pattern and added to the redundancy of genetic encoding.

Reproduction entailed choosing two different chromosomes from the population and recombining them into a single offspring genome. Moths that had been overlooked during predation trials were 2.5 times as likely to be chosen as the average detected individual. Within the detected group, moths were ranked in inverse order of the time the predator took to find them, and the highest-ranked individual had a 25% higher probability of being chosen than the lowest-ranked⁴⁴. To enable maintenance of integrated pattern features, recombination took place only between linkage groups, and the cross-over probability was determined by the combined values of the recombination regulators above and below the exchange point. Each offspring genome was subsequently subjected to a mutation process that randomly inverted individual bits with a probability of 0.003 (ref. 44). To standardize the impact of individual mutations, gene values were interpreted using Gray code, which minimizes the bit string distance between adjacent integer values⁴⁴.

Analytical measures

The primary dependent variables in this study were the crypticity and the phenotypic variability of the moths. To quantify crypticity, we calculated the frequency distribution of pixel intensities for each moth and computed the average unsigned difference between the moth and background distributions⁴⁵. In contrast to other studies of the mechanism of crypticity³⁵, we found no indication that differences between the moths and the background in the spatial frequency or distribution of patch sizes significantly affected detectability.

The crypticity index was subsequently validated by testing the consequences of direct manipulation of background resemblance, as well as by demonstrating that moths with a higher crypticity index were more difficult for the jays to find. Detection accuracy in blocks of 50 experimental trials declined significantly as a function of moth crypticity ($r^2 = 0.33$; $t(1) = 26.1$; $P < 0.0001$), and the response time for correct detections increased ($r^2 = 0.33$; $t(1) = 5.2$; $P < 0.0001$).

To quantify phenotypic variability, we applied a non-parametric partitioning algorithm⁴⁶ to the moth phenotypes in each generation, choosing a prototypical individual whose average distance from all others in phenotypic space (measured in terms of "taxonomic distance"⁴⁷) was minimal. This distance for each moth expressed the magnitude of its deviation from the prototype, and the average deviation for each population provided a relatively bias-free estimate of phenotypic variability.

Received 29 June; accepted 4 December 2001.

1. Dearn, J. M. in *Biology of Grasshoppers* (ed. Chapman, R. F. & Joern, A.) 517–549 (Wiley, New York, 1990).
2. Halkka, O. & Halkka, L. Population genetics of the polymorphic meadow spittlebug, *Philaenus spumarius* (L.). *Evol. Biol.* **24**, 149–191 (1990).
3. Edmunds, M. in *Insect Defenses* (ed. Evans, D. L. & Schmidt, J. O.) 3–21 (SUNY, Albany, New York, 1990).
4. Whiteley, D. A. A., Owen, D. F. & Smith, D. A. S. Massive polymorphism and natural selection in *Donacilla cornea* (Poli, 1791) (Bivalvia: Mesodesmatidae). *Biol. J. Linn. Soc.* **62**, 475–494 (1997).
5. Sargent, T. D. *Legion of Night: The Underwing Moths* (Univ. Massachusetts Press, Amherst, 1976).

6. Sargent, T. D. On the maintenance of stability in hindwing diversity among moths of the genus *Catocala* (Lepidoptera: Noctuidae). *Evolution* **32**, 424–434 (1978).
7. Barnes, W. & McDunnough, J. H. Illustrations of the North American species of the genus *Catocala*. *Mem. Am. Mus. Nat. Hist.* **3**, part 1, 1–47 (1918).
8. Owen, D. F. & Whiteley, D. Reflexive selection: Moment's hypothesis resurrected. *Oikos* **47**, 117–1120 (1986).
9. Common, I. F. B. A study of the ecology of the adult bogong moth, *Agrotis infusa* (Boisd.) (Lepidoptera: Noctuidae), with special reference to its behaviour during migration and aestivation. *Aust. J. Zool.* **2**, 223–263 (1954).
10. Pruess, K. P. Migration of the army cutworm, *Chorizagrotis auxiliaris* (Lepidoptera: Noctuidae). I. Evidence for a migration. *Ann. Entomol. Soc. Am.* **60**, 910–920 (1967).
11. Clarke, B. C. in *Taxonomy and Geography* (ed. Nichols, D.) 47–70 (Systematics Association, Oxford, 1962).
12. Allen, J. A. Frequency-dependent selection by predators. *Phil. Trans. R. Soc. Lond. B* **319**, 485–503 (1988).
13. Poulton, E. B. *The Colours of Animals* (Appleton, New York, 1890).
14. Tinbergen, L. The natural control of insects in pine woods. I. Factors influencing the intensity of predation by songbirds. *Arch. Néerlandaises Zool.* **13**, 265–343 (1960).
15. Bond, A. B. & Riley, D. A. Searching image in the pigeon: A test of three hypothetical mechanisms. *Ethology* **87**, 203–224 (1991).
16. Reid, P. J. & Shettleworth, S. J. Detection of cryptic prey: Search image or search rate? *J. Exp. Psychol. Anim. Behav. Process* **18**, 273–286 (1992).
17. Langley, C. M. Search images: Selective attention to specific visual features of prey. *J. Exp. Psychol. Anim. Behav. Process* **22**, 152–163 (1996).
18. Bond, A. B. & Kamil, A. C. Searching image in blue jays: Facilitation and interference in sequential priming. *Anim. Learn. Behav.* **27**, 461–471 (1999).
19. Bond, A. B. & Kamil, A. C. Apostatic selection by blue jays produces balanced polymorphism in virtual prey. *Nature* **395**, 594–596 (1998).
20. Rand, A. S. Predator–prey interactions and the evolution of aspect diversity. *Atlas Simp. Biota Amazônia* **5**, 73–83 (1967).
21. Ricklefs, R. E. & O'Rourke, K. Aspect diversity in moths: A temperate-tropical comparison. *Evolution* **29**, 313–324 (1975).
22. Allen, J. A. Reflexive selection is apostatic selection. *Oikos* **51**, 251–253 (1988).
23. Pietrewicz, A. T. & Kamil, A. C. Visual detection of cryptic prey by blue jays (*Cyanocitta cristata*). *Science* **195**, 580–582 (1977).
24. Pietrewicz, A. T. & Kamil, A. C. Search image formation in the blue jay (*Cyanocitta cristata*). *Science* **204**, 1332–1333 (1979).
25. Kono, H., Reid, P. J. & Kamil, A. C. The effect of background cuing on prey detection. *Anim. Behav.* **56**, 963–972 (1998).
26. Fleishman, L. J., McClintock, W. J., D'Erth, R. B., Brainard, D. M. & Endler, J. A. Colour perception and the use of video playback experiments in animal behaviour. *Anim. Behav.* **56**, 1035–1040 (1998).
27. Robinson, R. *Lepidopteran Genetics* (Pergamon, Oxford, 1971).
28. Nijhout, H. F. *The Development and Evolution of Butterfly Wing Patterns* (Smithsonian Institution, Washington DC, 1991).
29. Carroll, S. B. et al. Pattern formation and eyespot determination in butterfly wings. *Science* **265**, 109–114 (1994).
30. Brakefield, P. M. et al. Development, plasticity and evolution of butterfly eyespot patterns. *Nature* **384**, 236–242 (1996).
31. Guilford, T. & Dawkins, M. S. Search images not proven: A reappraisal of recent evidence. *Anim. Behav.* **35**, 1838–1845 (1987).
32. Endler, J. A. in *Behavioural Ecology* 3rd edn (eds Krebs, J. R. & Davies, N. B.) 169–196 (Blackwell Scientific, Oxford, 1991).
33. Bond, A. B. Visual search and selection of natural stimuli in the pigeon: The attention threshold hypothesis. *J. Exp. Psychol. Anim. Behav. Process* **9**, 292–306 (1983).
34. Wallace, A. R. *Darwinism: An Exposition of the Theory of Natural Selection with Some of its Applications* (MacMillan, London, 1891).
35. Endler, J. A. A predator's view of animal color patterns. *Evol. Biol.* **11**, 319–364 (1978).
36. Cott, H. B. *Adaptive Coloration in Animals* (Methuen, London, 1957).
37. Robinson, M. H. Defenses against visually hunting predators. *Evol. Biol.* **3**, 225–259 (1969).
38. Endler, J. A. Natural selection on color patterns in *Poecilia reticulata*. *Evolution* **34**, 76–91 (1980).
39. Ford, E. B. in *Insect Polymorphism: Symposia of the Royal Entomological Society of London* (ed. Kennedy, J. S.) Vol. 1, 11–19 (Royal Entomological Society, London, 1961).
40. Moment, G. B. Reflexive selection: a possible answer to an old puzzle. *Science* **136**, 262–263 (1962).
41. Nabours, R. K., Larson, I. & Hartwig, N. Inheritance of color patterns in the grouse locust *Acrydium arenosum* Burmeister (Tettigidae). *Genetics* **18**, 159–171 (1933).
42. Plaisted, K. C. & Mackintosh, N. J. Visual search for cryptic stimuli in pigeons: implications for the search image and search rate hypotheses. *Anim. Behav.* **50**, 1219–1232 (1995).
43. Darlington, C. D. & Mather, K. *The Elements of Genetics* (Macmillan, New York, 1950).
44. Bäck, T. *Evolutionary Algorithms in Theory and Practice* (Oxford Univ. Press, New York, 1996).
45. Endler, J. A. Progressive background matching in moths, and a quantitative measure of crypsis. *Biol. J. Linn. Soc.* **22**, 187–231 (1984).
46. Kaufman, L. & Rousseeuw, P. J. *Finding Groups in Data* (Wiley, New York, 1990).
47. Sneath, P. H. A. & Sokal, R. R. *Numerical Taxonomy* (Freeman, San Francisco, 1973).

Acknowledgements

We thank J. Allen, R. Balda, J. Endler, T. Getty, L. Harshman, S. Louda, D. Pilson and S. Shettleworth for their comments and suggestions. This research was supported by a grant from the National Science Foundation.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.B.B. (e-mail: abond@unl.edu) or A.C.K. (e-mail: akamil@unl.edu).

Testing the thermodynamic approach to granular matter with a numerical model of a decisive experiment

Hernán A. Makse* & Jorge Kurchan†

* Levich Institute and Physics Department, City College of the City University of New York, New York, New York 10031, USA

† PMMH, Ecole Supérieure de Physique et Chimie Industrielles, 10 rue Vauquelin, 75231 Paris, France

Edwards has proposed^{1–3} a thermodynamic description of dense, slowly flowing granular matter, in which the grains (the ‘atoms’ of the system) interact with inelastic forces and enduring contacts. In Edwards’ ensemble—one of the very few generalizations of standard statistical mechanics—thermodynamic quantities are computed as flat averages over configurations in which the grains are static or jammed, leading to a natural definition of configurational temperature. But the approach is not justified from first principles and hence, in the absence of explicit tests of its validity, has not been widely accepted. Here we report a numerical experiment involving a realistic model of slowly sheared granular matter; our results strongly support the thermodynamic description. Considering particles of different sizes in a slowly sheared dense granular system, we extract an effective temperature from a relation connecting their diffusivity and mobility. We then perform an explicit computation to show that the effective temperature measured from this relation coincides with the Edwards configurational temperature. Our approach, which is specifically conceived to be reproducible in the laboratory, may thus render the Edwards temperature accessible to experiments.

Consider a ‘tracer’ body of arbitrary shape immersed in a liquid in thermal equilibrium. As a consequence of the irregular bombardment by the particles of the surrounding liquid, the tracer performs a diffusive, fluctuating ‘brownian’ motion. The motion is unbiased, and for large times the average square of the displacement is given by $\langle |x(t) - x(0)|^2 \rangle = 2Dt$, where D is the diffusivity. On the other hand, if we pull gently on the tracer with a constant force f , the liquid responds with a viscous, dissipative force. The averaged displacement after a large time is $\langle |x(t) - x(0)| \rangle = f\chi t$, where χ is the mobility. Clearly, the same liquid molecules are responsible both for the fluctuations and for the dissipation. Although both D and χ strongly depend on the shape and size of the tracer, they turn out to be always related by the Einstein relation $D/\chi = T$ (a form of the fluctuation–dissipation theorem), where T is the temperature of the liquid. Conversely, if in a fluid of unknown properties we find that several tracers (of different sizes, for instance) having different diffusivities and mobilities yield the same ratio $D/\chi = T$, we may take this as a strong evidence for thermalization at temperature T .

The Einstein relation is strictly valid for an equilibrium thermal system. However, recent analytic schemes for out-of-equilibrium glassy systems have shown that the Einstein relation gives rise to a well-defined ‘effective temperature’ which is different from the bath temperature and governs the heat flow and the slow components of fluctuations and responses of all observables^{4,5}. The existence of an effective temperature with a thermodynamic meaning suggests a hidden form of ‘ergodicity’ for the slow modes of relaxation, which turns out to be closely related to Edwards’ statistical ideas¹ for systems undergoing slow compaction. Explicit checks of this approach have been made so far within the mean-field/mode-coupling models of the glass transition⁵, as well as for schematic finite-dimensional models of glassy systems^{6–10}.

On the experimental side, recently Nowak *et al.*¹¹ studied in detail the density fluctuations in a vibrated granular material, and

proposed that these fluctuations should reflect an underlying thermodynamics, much as they do in an ordinary thermal system. However, as far as the actual verification of the granular thermodynamics was concerned, the evidence they found was negative: we shall discuss below some possible reasons for this. On the other hand, although Edwards’ thermodynamics was originally formulated for granular matter, it has never been tested in realistic models of granular materials, characterized by the fact that energy is supplied by external driving (through, for instance, shear) and dissipated by inelastic collisions and slippage between the grains—a very different heat exchange mechanism from that of a thermal bath in a molecular system. In order to test the existence of an effective temperature with a thermodynamic meaning for dense, slowly moving granular matter, we perform, with a realistic numerical model, a diffusion-mobility experiment in conditions that can be reproduced in the laboratory. In the model, deformable spherical grains interact with one another via nonlinear elastic Hertz normal forces, nonlinear elastic and path-dependent Mindlin transverse forces, dissipative viscous damping force terms proportional to the relative normal and angular velocities, and via sliding friction with friction coefficient μ (see Methods ‘Microscopic model’; refs 12–14).

We perform molecular dynamics (MD) simulations (also known as discrete element methods, DEM)^{13–16} for a binary system of large and small spheres in a periodic three-dimensional cell. We apply a gentle simple shear flow at constant volume in the y -direction taking care that the gradient of the velocity along z is uniform (see Methods ‘MD simulation’). We focus our study on the slow shear rate, where the system is always close to jamming, but moves just barely enough to avoid stick-slip motion¹⁷. After a transient of the order of the inverse shear rate, we start measuring the spontaneous $\langle |x(t) - x(0)|^2 \rangle$ and force-induced displacements $\langle |x(t) - x(0)| \rangle / f$ along the x -direction for two types of particles. To measure

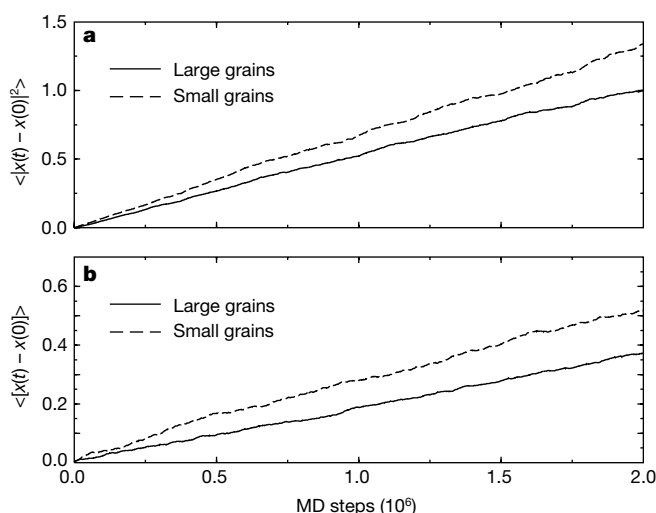


Figure 1 Diffusion (a) and response function (b) for the large and small particles in a sheared granular material measured perpendicular to the shear plane as a function of time in MD steps. Both quantities depend linearly on time at the late stages of the evolution. The response function is measured by applying a constant force f in the x -direction to each type of particle. Averages are taken over 30 realizations. The obtained diffusivities and mobilities depend on the particles size as expected. The slow frictional flow regime of interest in this study is accessed for shear rates $\dot{\gamma} \approx 10^{-3}$. The external pressure and mean solid fraction are large enough (~ 10 MPa and 0.66, respectively), and the average coordination number is high enough (typically ~ 7), that deformation and elasticity of the particles play the dominant role in the slow dynamics. The density of particles is uniform throughout the sample. The contacts between particles are enduring, and the internal stresses in the system are transmitted via a network of ‘force chains’^{24,25}.

$\langle |x(t) - x(0)| \rangle / f$ we apply an external small constant force ($f \approx 10^{-2}$ times the mean value of the contact forces) to the large and small particles separately and monitor their displacements as a function of time, averaging over all the small and large particles, respectively. The results are shown in Fig. 1. We notice that the diffusivities and the mobilities are different for the two type of particles, as expected. However, when we draw the parametric plot of $\langle |x(t) - x(0)|^2 \rangle$ versus $\langle |x(t) - x(0)| \rangle / f$ (Fig. 2) we find parallel straight lines for large timescales, implying an extended Einstein relation:

$$\langle |x(t) - x(0)|^2 \rangle = 2 T_{\text{eff}} \frac{\langle |x(t) - x(0)| \rangle}{f} \quad (1)$$

valid for both particles with the same 'effective temperature' T_{eff} for large timescales. This suggests that T_{eff} can be considered to be the temperature of the slow modes.

We also repeat the numerical experiment for a system of Hertz spheres without transverse forces ($\mu = 0$: experimental realizations of elastic spheres interacting with normal forces (contact and viscous forces) but without sliding friction are foams and compressed emulsions^{18–20}), and find that T_{eff} is well defined at long timescales for this case as well (see Fig. 2). Thus, our results suggest that the validity of an effective temperature for long-scale displacements (larger than a fraction of the particle size) holds in the presence of viscous forces between grains or even of a sliding threshold (Coulomb's law).

The existence of a single temperature is an instance of the 'zeroth law' of thermodynamics, for which we find positive evidence here, at variance with the experimental result of Nowak *et al.*¹¹. Three possible reasons for the apparent violation of the 'zeroth law' in

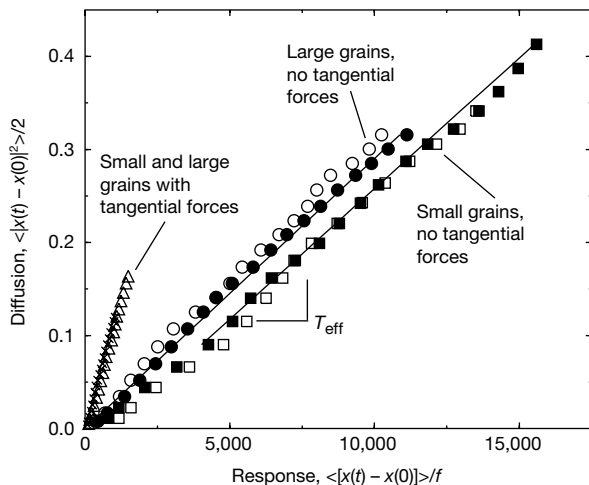


Figure 2 Parametric plot of diffusion versus response function for small and large grains interacting with tangential forces and without tangential forces (Coulomb frictionless). The fitting at long timescales shows the existence of a well defined effective temperature which is the same for small and large grains: $T_{\text{eff}} = 2.8 \times 10^{-5}$ for grains without transverse forces, and $T_{\text{eff}} = 1.2 \times 10^{-4}$ for grains with Mindlin transverse forces and Coulomb friction. These effective temperatures are about 10^{14} times kT at room temperature, as expected²⁶. We calculate the response function for several small external fields and find the same temperature, indicating that we are in the linear response regime. Plotted are results for a system without transverse forces using: $f = 1.7 \times 10^{-5}$ (small grains, open squares; large grains, open circles) and $f = 2.6 \times 10^{-5}$ (small grains, filled squares; large grains, filled circles). For a system with tangential forces and Coulomb friction we show the case $f = 6 \times 10^{-5}$ (small grains, triangles; large grains, crosses). For small timescales (fast rearrangements, not seen in this figure) this fluctuation-dissipation plot yields no evidence of a well-defined temperature. We expect this result, because the fast motion of the grains depends on the microscopic interactions dominated by inelastic collisions between grains. This two-relaxation model is analogous to that found for glasses, although in the case of glasses a well-defined temperature is also found for the fast rearrangements of the particles⁴.

their experiments are (1) the effect of an unknown height-dependent pressure (as pointed out by the authors), (2) a rather high tapping amplitude (a well-defined T_{eff} is only to be expected in the high compaction limit, ref. 5), and (3) the fact that the density fluctuations considered are integrated over all frequencies, thus including, unlike the present work, also 'fast' relaxations.

Next, we treat the question whether it is possible to relate the effective temperature obtained above to the thermodynamic construction of out-of-equilibrium systems proposed by Edwards. While in the Boltzmann–Gibbs construction of statistical mechanics one assumes that the physical quantities are obtained as averages over all possible configurations, the Edwards ensemble consists of only the jammed (blocked or static) configurations at the appropriate energy and volume. The strong 'ergodic' hypothesis is that all jammed configurations of a given volume and energy can be taken to have equal statistical weights. This formulation leads to a configurational entropy $S_{\text{Edw}}(E, V) = \ln \Omega_{\text{Edw}}(E, V)$, where Ω_{Edw} is the number of jammed configurations, and the corresponding configurational temperature $T_{\text{Edw}}^{-1} = \partial S_{\text{Edw}} / \partial E$ and compactivity $X_{\text{Edw}}^{-1} = \partial S_{\text{Edw}} / \partial V$.

In order to calculate T_{Edw} and compare it with the obtained T_{eff} , we need to count the number of jammed configurations at a given energy and volume (for this calculation we concentrate on the case without tangential forces and sliding friction, in order to avoid path dependency which would lead to an ambiguity in the definition of jammed configurations—see below). Counting directly all jammed configurations is impossible, except for small systems. To do it in practice, we resort here to an indirect 'auxiliary model' method⁷ suitably modified to the case of deformable grains. It consists of computing, using any standard method (Monte Carlo, MD, and so on), the equilibrium properties of the granular system

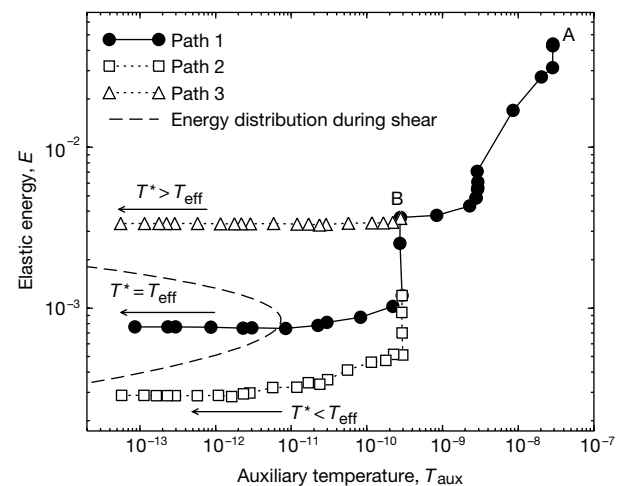


Figure 3 Annealing procedure to calculate T_{Edw} at different elastic compressional energies. We plot the elastic energy versus T_{aux} during the annealing, together with the distribution of elastic energies obtained during shear (dashed curve, mean value $\langle E \rangle = 8.4 \times 10^{-4}$). We equilibrate the system for 40×10^6 iterations at A: ($T^* = 3.4 \times 10^{-2}$, $T_{\text{aux}} = 3 \times 10^{-8}$). We then anneal slowly both temperatures until B: ($T^* = 3.4 \times 10^{-4}$, $T_{\text{aux}} = 3 \times 10^{-10}$), where we split the trajectory into three paths in the (T^* , T_{aux}) plane. Path 1: we anneal $T_{\text{aux}} \rightarrow 0$ and $T^* \rightarrow 2.8 \times 10^{-5}$ which corresponds to T_{eff} obtained during shear (Fig. 2). Path 2: we anneal $T_{\text{aux}} \rightarrow 0$ and $T^* \rightarrow 3.4 \times 10^{-6}$. Path 3: we anneal $T_{\text{aux}} \rightarrow 0$ but keep $T^* = 3.4 \times 10^{-4}$ constant. When we set $T^* = T_{\text{eff}}$ (Path 1), the final elastic compressional energy value when $T_{\text{aux}} \rightarrow 0$ falls inside the distribution of energies obtained, and it is very close to the mean value of the elastic energy during shear ($\langle E \rangle$). This proves that $T_{\text{eff}} = T_{\text{Edw}}$ under the numerical accuracy of the simulations. For other values of $T^* \neq T_{\text{eff}}$, the final E falls out of the distribution obtained during shear (paths 2 and 3). We also follow different trajectories (not shown in the figure) to $T^* \rightarrow 2.8 \times 10^{-5}$, $T_{\text{aux}} \rightarrow 0$ and find the same results, indicating that our procedure is independent of the annealing path.

in the periodic cell with the modified partition function $\Sigma \exp[-E/T^* - E_{\text{jammed}}/T_{\text{aux}}]$, where E is the elastic compressional potential energy leading to the Hertz contact force, and $E_{\text{jammed}} \approx \Sigma_a |F_a|^2$, with F_a the total contact force exerted on particle a by its neighbours. Thus, the dynamical system studied previously via the Einstein relation is now augmented by a temperature T^* setting the mean elastic energy per grain, plus an auxiliary temperature, T_{aux} , which relates to the (artificial) term E_{jammed} which selects, in the limit $T_{\text{aux}} \rightarrow 0$, the configurations where the grains are jammed ($E_{\text{jammed}} \rightarrow 0$).

It can be shown, although it may not be immediately obvious, that the following protocol yields the correct Edwards temperature at energy E . We start by equilibrating the system at high temperatures (T_{aux} and $T^* \approx \infty$), and anneal slowly the value of T_{aux} to zero while tuning T^* so as to achieve a given final elastic potential energy E . At the end of the procedure ($T_{\text{aux}} \rightarrow 0$), we read Edwards' temperature as $T_{\text{Edw}}(E) = T^*(E)$. This simulation is performed with the same system parameters as in the previous dynamical simulations within the shearing cell, but without viscous internal dissipation and external shear. We note that the procedure is a calculational trick, and it does not correspond to any real experimental protocol.

In Fig. 3 we plot the elastic compressional energy as a function of T_{aux} for three annealing protocols using different values of T^* ; we also show the distribution of compressional elastic energies obtained during the shearing experiments. The y -intercept of each annealing curve gives, as mentioned above, the value of E at $T_{\text{Edw}} = T^*$. Only when we set T^* equal to T_{eff} (as obtained through the Einstein relation in the shearing cell), do we find that the final E coincides, within the accuracy of the simulations, with the mean elastic compressional potential energy of the system under shear (Fig. 3). This shows the agreement between Edwards' T_{Edw} and the effective temperature T_{eff} .

We conclude with some remarks. (1) Because the jammed configurations are the same whatever the inter-grain dissipation coefficient, Edwards' ensemble (and hence its temperature) is insensitive to viscous dissipation, as long as we are in the slow flow regime of interest in this study. (2) But tangential forces and sliding friction may or may not block certain configurations, depending on how they are accessed: the ensemble of jammed configurations is then ill-defined. We have not tried to construct a suitable ensemble for this case, but content ourselves with the observation that T_{eff} is also in this case independent of the particle size (Fig. 2)—our results suggest that thermodynamic concepts still apply, but the relevant ensemble for frictional systems necessarily goes beyond Edwards' construction as it stands. (3) We have tested the validity of the thermodynamics in an ideal homogeneous system with periodic boundary conditions by explicitly avoiding structural features of dense granular flows such as inhomogeneities and shear bands (by imposing a uniform shear rate), segregation of the species or long-range order^{21,22}. Thus, the experimental test of our computational results may be complicated by, for instance, the formation of shear bands which tend to be present in physical systems. Even though it remains to be seen whether the thermodynamic picture may account for these highly nonlinear and dissipative effects, our ideal system may prove to be useful in deriving constitutive relations to be used in macroscopic theories of slow granular flows.

We have performed numerically a diffusion-mobility experiment with a dense, slowly sheared granular system specially designed to be a 'dress rehearsal' for the real laboratory experiment. The non-dependence of T_{eff} on the tracer size provides strong evidence for an underlying thermodynamics. We have then independently computed the configurational temperature T_{Edw} based on the entropy of jammed configurations and verified that it coincides with T_{eff} —thus supporting Edwards' statistical mechanical ideas. This last step cannot be performed in the laboratory, so the numerical simulation

provides the missing link between thermodynamic ideas and diffusion-mobility checks. □

Methods

Microscopic model

In the simulations, two grains in contact interact via a normal Hertz force, F_n , and a tangential Mindlin force F_t (ref. 12). For two spherical grains with radii R_1 and R_2 : $F_n = \frac{2}{3} k_n R^{1/2} w^{3/2}$, $\Delta F_t = k_t (Rw)^{1/2} \Delta s$. Here $R = 2R_1 R_2 / (R_1 + R_2)$, the normal overlap is $w = (1/2)(R_1 + R_2) - |\mathbf{x}_1 - \mathbf{x}_2| > 0$, where $\mathbf{x}_1, \mathbf{x}_2$ are the positions of the grain centres. The normal force acts only in compression, $F_n = 0$ when $w < 0$. The variable s is defined such that the relative shear displacement between the two grain centres is $2s$. The prefactors $k_n = 4G/(1 - \nu)$ and $k_t = 8G/(2 - \nu)$ are defined in terms of the shear modulus G and the Poisson's ratio ν of the material from which the grains are made (typically $G = 29$ GPa and $\nu = 0.2$, for glass beads). When F_t exceeds the Coulomb threshold, μF_n , the grains slide and $F_t = \mu F_n$, where μ is the friction coefficient between the spheres (typically $\mu = 0.3$). We assume a distribution of grain radii in which $R_1 = 0.105$ mm for half the grains and $R_2 = 0.075$ mm for the other half. The observables are measured in reduced units: length in units of R , force in units of GR , time in units of $\sqrt{\rho R^2/G}$, where ρ is the density of the particles.

MD simulations

We perform MD simulations for a system of 200 particles. Our calculations begin with a numerical protocol designed to mimic the experimental procedure used to prepare dense packed granular materials¹⁵. The simulations begin with a gas of spherical particles located at random positions in a periodically repeated cubic cell of side L . The system is then compressed and extended slowly until a specified value of the pressure and volume fraction—above the random close packing fraction—is achieved at static equilibrium. We then apply a gentle shear in the y -direction at constant volume by moving the periodic images at the top and bottom of the cell with velocities $\dot{\gamma}L/2$, where $\dot{\gamma}$ is the shear rate (Lees–Edwards boundary conditions) and using a suitable modified set of equations (see ch. 8 of ref. 16, and ref. 23) to generate a linear uniform shear velocity profile along the z -direction. Periodic boundary conditions are enforced in the x -direction and in the y -direction of the flow. We checked that shear-induced segregation is absent at the timescales of our simulations.

Received 10 July; accepted 20 December 2001.

- Edwards, S. F. in *Granular Matter: An Interdisciplinary Approach* (ed. Mehta, A.) 121–140 (Springer, New York, 1994).
- Edwards, S. F. in *Disorder in Condensed Matter Physics* (eds Blackman, J. & Taguena, J.) 147–154 (Oxford Univ. Press, Oxford, 1991).
- Mehta, A. & Edwards, S. F. Statistical mechanics of powder mixtures. *Physica A* **157**, 1091–1097 (1989).
- Cugliandolo, L. F., Kurchan, J. & Peliti, L. Energy flow, partial equilibration, and effective temperatures in systems with slow dynamics. *Phys. Rev. E* **55**, 3898–3914 (1997).
- Kurchan, J. in *Jamming and Rheology: Constrained Dynamics on Microscopic and Macroscopic Scales* (eds Liu, A. & Nagel, S. R.) 72–79 (Taylor and Francis, London, 2001).
- Nicodemi, M. Dynamical response functions in models of vibrated granular media. *Phys. Rev. Lett.* **82**, 3734–3737 (1999).
- Barrat, A., Kurchan, J., Loreto, V. & Sellitto, M. Edwards measures for powders and glasses. *Phys. Rev. Lett.* **85**, 5034–5037 (2000).
- Javier Brey, J., Prados, A. & Sanchez-Rey, B. Thermodynamic description in a simple model for granular compaction. *Physica A* **275**, 310–324 (2000).
- Lefevre, A. & Dean, D. S. Tapping spin-glasses and ferromagnets on random graphs. *Phys. Rev. Lett.* **86**, 5639–5642 (2001).
- Sciortino, F. & Tartaglia, P. Extension of the fluctuation-dissipation theorem to the physical aging of a model glass-forming liquid. *Phys. Rev. Lett.* **86**, 107–110 (2001).
- Nowak, E. R., Knight, J. B., Ben-Naim, E., Jaeger, H. M. & Nagel, S. R. Density fluctuations in vibrated granular materials. *Phys. Rev. E* **57**, 1971–1974 (1998).
- Johnson, K. L. *Contact Mechanics* (Cambridge Univ. Press, Cambridge, 1985).
- Cundall, P. A. & Strack, O. D. L. A Discrete numerical model for granular assemblies. *Géotechnique* **29**, 47–65 (1979).
- Schäfer, J., Dippel, S. & Wolf, D. E. Force schemes in simulations of granular materials. *J. Phys. I (France)* **6**, 5–20 (1996).
- Makse, H. A., Johnson, D. L. & Schwartz, L. M. Packing of compressible granular materials. *Phys. Rev. Lett.* **84**, 4160–4163 (2000).
- Allen, M. P. & Tildesley, D. J. *Computer Simulations of Liquids* (Clarendon, Oxford, 1987).
- Savage, S. B. The mechanics of rapid granular flows. *Adv. Appl. Mech.* **24**, 289–365 (1994).
- Durian, D. J. Foam mechanics at the bubble scale. *Phys. Rev. Lett.* **75**, 4780–4783 (1995).
- Lacasse, M.-D., Grest, G. S., Levine, D., Mason, T. G. & Weitz, D. A. Model for the elasticity of compressed emulsions. *Phys. Rev. Lett.* **76**, 3448–3451 (1996).
- Langer, S. A. & Liu, A. J. Sheared foam as a supercooled liquid? *Europhys. Lett.* **49**, 68–73 (2000).
- Herrmann, H. J., Hovi, J. P. & Luding, S. (eds) *Physics of Dry Granular Matter* (Kluwer, Dordrecht, 1998).
- Ottino, J. M. & Khakhar, D. V. Mixing and segregation of granular materials. *Annu. Rev. Fluid Mech.* **32**, 55–91 (2000).
- Hoover, W. G. et al. Lennard-Jones triple-point bulk and shear viscosities. Green-Kubo theory, Hamiltonian mechanics, and non-equilibrium molecular dynamics. *Phys. Rev. A* **22**, 1690–1697 (1980).
- Drescher, A. & de Josselin de Jong, G. Photoelastic verification of a numerical model for the flow of a granular material. *J. Mech. Phys. Solids* **20**, 337–351 (1972).
- Liu, C.-H. et al. Force fluctuations in bead packs. *Science* **269**, 513–515 (1995).
- Jaeger, H. M., Nagel, S. R. & Behringer, R. P. Granular solids, liquids, and gases. *Rev. Mod. Phys.* **68**, 1259–1273 (1996).

Acknowledgements

This work was partially supported by The Petroleum Research Fund.

Correspondence and requests for materials should be addressed to H.A.M. (e-mail: makse@mailaps.org).

Growth of nanowire superlattice structures for nanoscale photonics and electronics

Mark S. Gudiksen^{*†}, Lincoln J. Lauhon^{*†}, Jianfang Wang^{*}, David C. Smith^{*‡} & Charles M. Lieber^{*§}

^{*} Department of Chemistry and Chemical Biology, [§] Division of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA

[†] These authors contributed equally to this work

The assembly of semiconductor nanowires and carbon nanotubes into nanoscale devices and circuits could enable diverse applications in nanoelectronics and photonics¹. Individual semiconducting nanowires have already been configured as field-effect transistors², photodetectors³ and bio/chemical sensors⁴. More sophisticated light-emitting diodes⁵ (LEDs) and complementary and diode logic^{6–8} devices have been realized using both n- and p-type semiconducting nanowires or nanotubes. The n- and p-type materials have been incorporated in these latter devices either by crossing p- and n-type nanowires^{2,5,6,9} or by lithographically defining distinct p- and n-type regions in nanotubes^{8,10}, although both strategies limit device complexity. In the planar semiconductor industry, intricate n- and p-type and more generally compositionally modulated (that is, superlattice) structures are used to enable versatile electronic and photonic functions. Here we demonstrate the synthesis of semiconductor nanowire superlattices from group III–V and group IV materials. (The superlattices are created within the nanowires by repeated modulation of the vapour-phase semiconductor reactants during growth of the wires.) Compositionally modulated superlattices consisting of 2 to 21 layers of GaAs and GaP have been prepared. Furthermore, n-Si/p-Si and n-InP/p-InP modulation doped nanowires have been synthesized. Single-nanowire photoluminescence, electrical transport and electroluminescence measurements show the unique photonic and electronic properties of these nanowire superlattices, and suggest potential applications ranging from nano-barcodes to polarized nanoscale LEDs.

Our approach to superlattice growth (Fig. 1) exploits recent developments in metal-catalysed nanowire synthesis, which have shown that nearly monodisperse metal nanoclusters can be used to control the diameter^{11,12} and (through growth time) the length¹³ of group III–V and group IV semiconductor nanowires by way of a vapour–liquid–solid growth process¹⁴. We introduce the vapour-phase semiconductor reactants required for nanowire growth by either laser ablation of solid targets^{15,16} or vapour-phase molecular species^{12,17}. To create a single junction within the nanowire, the addition of the first reactant is stopped during growth, and then a second reactant is introduced for the remainder of the synthesis (Fig. 1b); repeated modulation of the reactants during growth produces nanowire superlattices (Fig. 1c). In principle, this

approach can be successfully implemented if a nanocluster catalyst suitable for growth of the different superlattice components under similar conditions is found; our previous studies¹⁶ suggest that Au nanoclusters meet this requirement for a wide range of III–V and IV materials.

Gallium arsenide (GaAs)/gallium phosphide (GaP) superlattices have been grown by laser-assisted catalytic growth using GaAs and GaP targets. Figure 2 shows transmission electron microscopy (TEM) images of the products of this synthesis. It is relatively straightforward to focus on the junction area as the nanowire lengths can be controlled directly by growth times¹³. High-resolution TEM images of a typical GaAs/GaP junction region (Fig. 2a) exhibit a crystalline nanowire core without obvious defects, and show that the nanowire axes lies along the $\langle 111 \rangle$ direction, in agreement with previous studies of single-component systems^{11,13,15,16}. Two-dimensional Fourier transforms calculated from high-resolution images containing the junction region (Fig. 2a, inset) show pairs of reciprocal lattice peaks along the different lattice directions, while such transforms calculated from the regions above and below the junction (not shown) exhibit only single reciprocal lattice peaks. Analysis of these peak data yield lattice constants, indexed to the zinc blende structures of GaP and GaAs, of 0.5474 ± 0.0073 nm and 0.5668 ± 0.0085 nm, and are in good agreement with the values for both GaP (0.5451 nm) and GaAs (0.5653 nm), respectively.

The TEM structural data suggest that the GaP/GaAs junctions could be abrupt, and thus we have carried out local elemental mapping of the heterojunction by energy dispersive X-ray spectroscopy (EDS) to address composition variation across the junction (Fig. 2b–e). These elemental maps show that Ga is uniformly distributed along the length of the nanowire, while P (Fig. 2d) and As (Fig. 2e) appear localized in the GaP and GaAs portions of the nanowire heterostructure, respectively. Quantitative analysis of the P and As composition variation (Fig. 2f) shows, however, that the junction is not atomically abrupt, but rather makes the transition between GaP and GaAs phases over a length scale of 15–20 nm. This length scale is reasonable considering that the ~20-nm-diameter Au catalyst must re-alloy with GaP after initial GaAs growth. The observed composition variation has several potentially important implications. First, composition variation at the interface can relieve strain, and may enable the defect-free junctions and superlattices that we observe in this system—which has a relatively large lattice mismatch. We note, however, that simple estimates of the length between dislocations (M.S.G. and C.M.L., unpublished results) suggest that defect-free, atomically abrupt interfaces may be possible in wires of diameter less than 20 nm. Second, there are photonic and electronic applications where abrupt interfaces are important. The observed composition variation could be substantially reduced in smaller-diameter nanowires; that is, a 5-nm-

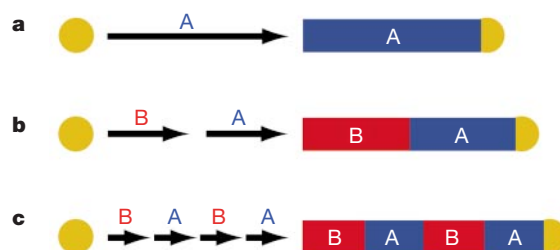


Figure 1 Synthesis of nanowire superlattices. **a**, A nanocluster catalyst (shown gold) nucleates and directs one-dimensional semiconductor nanowire (blue) growth with the catalyst remaining at the terminus of the nanowire. **b**, Upon completion of the first growth step, a different material (red) can be grown from the end of the nanowire. **c**, Repetition of steps **a** and **b** leads to a compositional superlattice within a single nanowire.

[‡]Present address: Department of Physics and Astronomy, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

diameter nanowire superlattice should have variations of <5 nm across the junction interfaces. Alternatively, it should be possible to use different nanocluster catalysts or variations in the growth temperature when reactants are switched to obtain sharper interfaces.

We find that this approach can produce compositionally modulated nanowire superlattices in which the number of periods and repeat spacing can be readily varied during growth. TEM images of a six-period structure corresponding to a $(\text{GaP}/\text{GaAs})_3$ superlattice (Fig. 3a) show that the 20-nm-diameter nanowires are highly uniform over their $\sim 3\text{-}\mu\text{m}$ lengths. Spatially resolved EDS measurements of composition (Fig. 3b) further demonstrate that the P and As regions are distinct from one another, and that there is minimal cross-contamination. Moreover, these data show that each GaP and GaAs nanowire segment has a length of about 500 nm, and are thus consistent with the equal growth times used for each segment, but also show that growth rates remain relatively constant during the entire nanowire synthesis.

GaAs/GaP nanowire superlattices are an attractive system to explore for nano-phonic applications because GaAs is a direct bandgap semiconductor, while GaP has an indirect gap. Indeed, photoluminescence imaging of individual nanowires from the $(\text{GaP}/\text{GaAs})_3$ superlattice sample described above shows that these nanowires exhibit an emission pattern of three spots separated by dark regions (Fig. 3c). This pattern is consistent with emission

originating from the three GaAs regions, separated by dark GaP regions that act as optical 'spacers'. Control experiments on individual samples of pure GaAs and GaP nanowires confirm that strong luminescence is obtained from GaAs but not GaP, as expected. The GaAs regions also exhibit a strong polarization dependence, emitting when the excitation is polarized parallel ($\parallel$) to the nanowire axis and appearing dark when the polarization is perpendicular ($\perp$) to the nanowire axis (Fig. 3c, inset). In addition, experiments (M.S.G., D.C.S. and C.M.L., unpublished results) show that the emission from the superlattice structures is also highly polarized along the wire axis. These results are consistent with recently reported photoluminescence spectroscopy of single InP nanowires³, and could enable several unique applications (see below).

We have also used single-nanowire photoluminescence studies to assess the level of control possible in nanowire superlattice growth by our approach. Specifically, we find that systematic variations in the growth time produces nanowire superlattices with well-defined changes in periodicity. The photoluminescence image of an 11-layer superlattice in which the length of the GaP regions was doubled each layer while maintaining a constant GaAs period (Fig. 3d) shows clearly that the separation between emitting GaAs regions is doubling along the length of the nanowire. In addition, the power and promise of this approach can be seen clearly in photoluminescence spectra of 21-layer GaP/GaAs superlattices (Fig. 3e) consisting of a short 4-period (GaP/GaAs) repeat, followed by 3 longer GaP

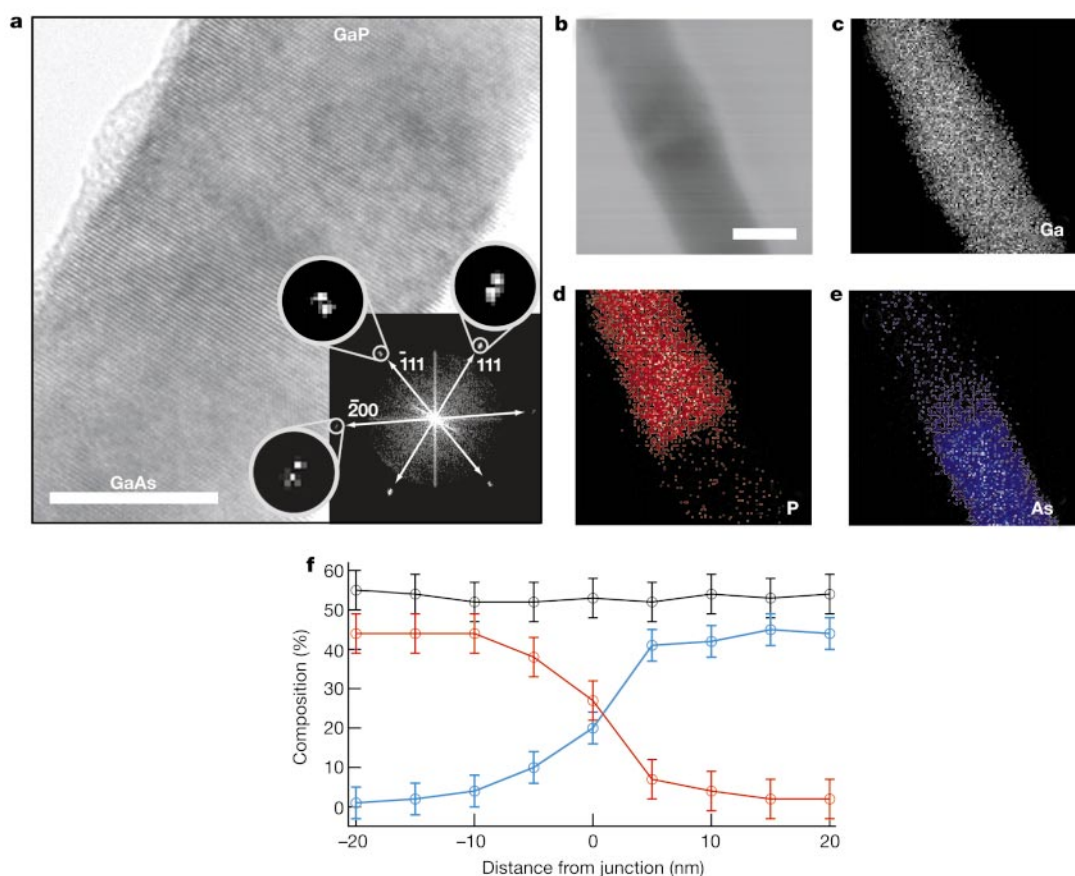


Figure 2 GaAs/GaP nanowire junctions. **a**, High-resolution TEM of a GaAs/GaP junction grown from a 20-nm gold nanocluster catalyst. Scale bar, 10 nm. Inset, two-dimensional Fourier transforms of the entire image show a splitting of the reciprocal lattice peaks along the $\langle 111 \rangle$, $\langle \bar{1}\bar{1}\bar{1} \rangle$ and $\langle 200 \rangle$ lattice directions in the $[022]$ zone axis, corresponding to the lattice constants for GaAs and GaP (see text). The presence of the heterojunction was confirmed by EDS analysis above and below the junction region (not shown). **b**, TEM image of another junction. Scale bar, 20 nm. **c**, **d**, **e**, Elemental mapping of the Ga (shown

grey), P (red) and As (blue) content of the junction shown in **b**. A scanning TEM was used to take an elemental 'image' of the junction. **f**, Line profiles of the composition through the junction region, showing the change in composition as a function of the distance. The slightly higher Ga (shown black) signal relative to the P (red) and As (blue) signals may be due to uncertainties in the detector calibration or the presence of an amorphous gallium oxide layer around the crystalline nanowire core.

spacer repeats, and ending in a relatively short 4-period (GaAs/GaP) repeat. We consider that in their present form these nanowire superlattice structures could be exploited as optical nanobarcodes¹⁸, which could be useful as labels for imaging. Moreover, the wide range of group III–V and II–VI nanowires that have been

demonstrated previously¹⁶ suggests that it should be possible to encode additional information through variations in the colour of the emitting region using multi-component superlattices. Using materials with a large dielectric contrast might also enable the creation of one-dimensional waveguides¹⁹ with built-in photonic bandgaps, or of cavities for nanowire lasers²⁰.

To assess further the generality of this methodology, we fabricated p–n junctions within individual silicon nanowires by Au-nanocluster-catalysed chemical vapour deposition and dopant modulation^{12,17}. These nanowire p–n junctions were characterized at the single nanowire level by a variety of electrical measurements (Fig. 4) because EDS is insufficiently sensitive to characterize

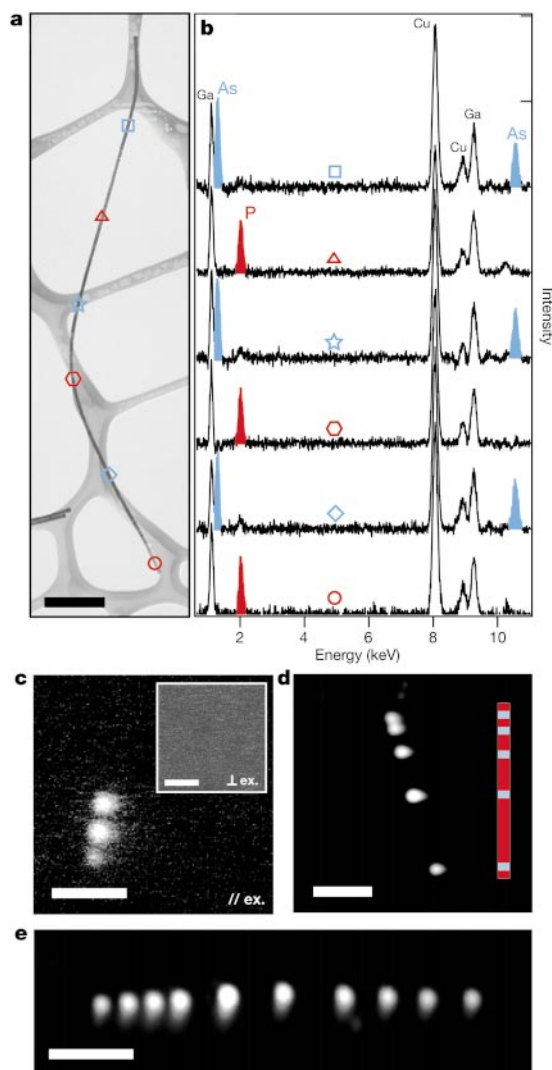


Figure 3 Nanowire superlattice structures. **a**, TEM image of a ~20-nm-diameter (GaP/GaAs)₃ nanowire superlattice. The background mesh is from the holey carbon film, on which the nanowire was deposited for imaging. Scale bar, 300 nm. **b**, Elemental profile of the superlattice along the nanowire length measured by EDS analysis. The symbols in **a** show the location of each EDS spectrum along the nanowire and the colour of the symbol indicates GaP (red) and GaAs (blue) regions. The P K α peak (2.015 keV) is shown in red and the As K α (10.543 keV) and L α (1.282 keV) peaks in blue for clarity. The spectra show clearly a distinct, periodic modulation of the nanowire composition along its entire length, with three uniform periods of GaP, separated by three uniform periods of GaAs. **c**, Photoluminescence image (// excitation) of a nanowire from the same sample as shown in **a** and **b**. The three bright regions correspond to the three GaAs (direct bandgap) regions, while the dark segments are from the GaP (indirect bandgap) regions. Inset, no photoluminescence is observed above background for perpendicular ($\perp$) excitation due to the dielectric contrast between the nanowire and its surroundings³. Scale bar, 5 μ m. **d**, Photoluminescence image of a 40-nm-diameter GaP(5)/GaAs(5)/GaP(5)/GaAs(5)/GaP(10)/GaAs(5)/GaP(20)/GaAs(5)/GaP(40)/GaAs(5)/GaP(5) superlattice; the numbers in parentheses correspond to the growth times in seconds for each layer. Inset, diagram showing the relative lengths of GaAs (blue) and GaP (red) layers. Scale bar, 5 μ m. **e**, Photoluminescence image of a 21-layer superlattice, (GaP/GaAs)₁₀GaP, showing a group of four equally spaced spots on the left, two in the middle with larger gaps, and another set of four with equal spacing on the end. The superlattice is ~25 μ m in length.

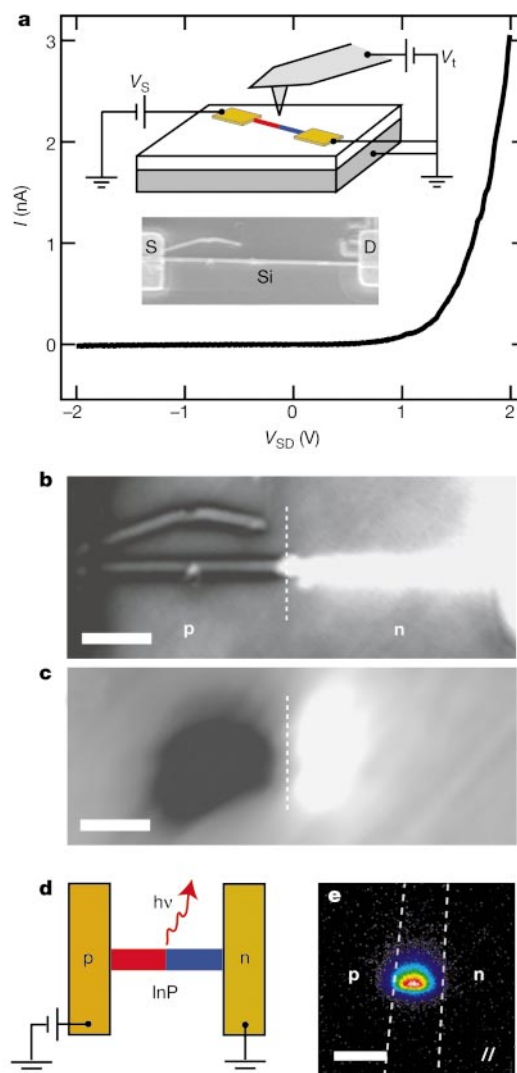


Figure 4 Modulation-doped nanowires. **a**, Inset: a diagram illustrating single-nanowire electrical characterization by transport and probe microscopy, and a scanning electron micrograph of the silicon nanowire device with source (S) and drain (D) electrodes indicated. Main panel, I versus V_{SD} for the silicon nanowire p–n junction. **b**, EFM phase image of the nanowire diode under reverse bias with tip bias (V_t) at +3 V and the drain (right) at +2 V. The signal is proportional to the square of the tip–surface potential difference, and shows an abrupt drop in the middle of the wire at the junction. **c**, SGM image showing the source–drain current as the tip (at +10 V) is scanned across the device. With the drain biased at –2 V ($V_{SD} = +2$ V), bright (dark) regions correspond to an increase (decrease) in the positive quantity I_{SD} . Vertical dashed white lines indicate the junction in **b** and **c**. Scale bars, 500 nm. **d**, Schematic of a InP nanowire LED. **e**, Polarized emission from the LED along the nanowire axis. Dashed white lines indicate the edges of the electrodes in **d**, and were determined from a white light image. No electroluminescence was detected with perpendicular polarization. Scale bar, 3 μ m.

dopant profiles. Current (I) versus voltage (V_{SD}) measurements showed rectifying behaviour consistent with the presence of an intra-nanowire p–n junction (Fig. 4a). To establish that the current rectification was due to intra-nanowire p–n junctions, we characterized the local nanowire potential and gate response by electrostatic force microscopy (EFM) and scanned gate microscopy (SGM), respectively²¹. An EFM image of a typical p–n junction in reverse bias showed that the entire voltage drop occurs at the p–n junction itself (Fig. 4b); EFM measurements showed no potential drop at the contact regions under forward or reverse bias (not shown), ruling out the contact/nanowire interface as the source of rectification in the I – V_{SD} behaviour. In addition, SGM images recorded with the nanowire device in forward bias and the scanned tip–gate positive (Fig. 4c) show enhanced conduction to the right of the junction, indicating an n-type region, and reduced conduction to the left of the junction, indicating depletion of a p-type region. The abrupt change in majority carrier type coincides with the location of the intra-nanowire junction determined by EFM, and thus further confirms that the diode behaviour results from well controlled dopant modulation.

We believe that the direct and controlled growth of nanowire p–n junctions represents an advance over previous work^{8,10,22}. The ability to synthesize modulation-doped nanowire superlattices opens up new opportunities, ranging from ultra-sensitive biological and chemical detectors to bipolar transistors and highly integrated logic gates for nanoelectronics. Moreover, the direct growth of modulation-doped nanowires eliminates the lithographic steps used to create doped nanotube p–n junctions^{8,10}, and thus facilitates the bottom-up assembly of complex functional structures when combined with recent advances in the directed *en masse* organization of nanowire structures^{5,9}.

Last, the use of single InP nanowire p–n junctions as nanoscale LEDs has been investigated (Fig. 4d). I – V_{SD} measurements of InP nanowire p–n structures exhibit rectification similar to that described above for the silicon nanowire p–n junctions. In forward bias, individual InP nanowire devices exhibit light emission from p–n junctions that is both highly polarized and blue-shifted due to the one-dimensional structure and radial quantum confinement, respectively (Fig. 4e)³. The efficiency of these intra-nanowire LEDs is ~0.1%, although it can be increased substantially. By defining a quantum dot heterostructure within a p–n diode during nanowire synthesis, it should be possible to engineer an electrically driven single-photon source with well-defined polarization. Such a nanowire device could be extremely useful in quantum cryptography and information processing²³. More generally, we believe that the present results open up many opportunities in nanoscale photonics and electronics, ranging from the relatively simple nanoscale emitters and complementary logic, which can be obtained from single nanowire p–n junctions, to complex periodic superlattices that may enable applications such as nanowire injection lasers and ‘engineered’ one-dimensional electron waveguides. □

Methods

Nanowire synthesis

Nanowires were synthesized either via laser-assisted catalytic growth (GaAs, GaP and InP) or chemical vapour deposition (Si), using Au nanoclusters to direct the growth. Au nanoclusters were deposited onto oxidized silicon substrates and then placed in the reactor furnace. For nanowires produced by laser-assisted catalytic growth, solid targets of GaAs, GaP and InP were ablated using either a pulsed ArF excimer or Nd:YAG lasers, and growth was carried out at 700–850 °C in an argon flow of 100 cm³ STP per minute at 100 torr. A pause of ~45 s in the ablation was made between each layer in a given superlattice. Silicon nanowires were grown by chemical vapour deposition at 450 °C using silane (3 cm³ STP per minute) and either 100 p.p.m. diborane (p) or phosphine (n) in helium (18 cm³ STP per minute) as dopants. The furnace was evacuated before switching dopants.

The resulting nanowires were sonicated briefly in ethanol and deposited onto copper grids for TEM analysis. The high-resolution TEM images and EDS spectra from nanowire superlattices were collected on a JEOL 2010F microscope. The elemental mapping of the single junction was conducted on a VG HB603 STEM.

Photoluminescence measurements

Single-nanowire photoluminescence images were obtained using a purpose-built, far-field epifluorescence microscope. Excitation light (488 nm) was focused by an objective (NA = 0.7) to a ~30-μm-diameter spot at ~1.0 kW cm⁻² on the quartz substrate with dispersed nanowires. A $\lambda/2$ wave plate was used to rotate the polarization of the excitation light, and the resulting photoluminescence was focused and imaged on a liquid-nitrogen-cooled charge coupled device. The emission polarization was measured using a Glan–Thompson polarizer.

Electrical measurements

Nanowires dispersed in ethanol were deposited onto silicon substrates (600-nm oxide), and electrical contacts were defined using electron beam lithography. Ti/Au contacts were used for Si nanowires, and were annealed at 400 °C following deposition. InP LED contacts were fabricated by a two-step process in which the first contact (n-type) was made using Ge/Au or Ni/In/Au and the second (p-type) was made using Zn/Au. The contacts were annealed at 300–350 °C following deposition.

Scanned probe microscopy

A Digital Instruments Nanoscope III with extender module was used for the EFM and SGM measurements. Force modulation etched silicon probe (FESP) tips coated with 5 nm Cr/45 nm Au were used for imaging. For EFM, the nanoscope was operated in lift mode with a lift height of 60 nm and a scan rate of 0.5 Hz.

Received 24 December 2001; accepted 11 January 2002.

- Lieber, C. M. The incredible shrinking circuit. *Sci. Am.* **285**, 58–64 (2001).
- Cui, Y. & Lieber, C. M. Functional nanoscale electronic devices assembled using silicon nanowire building blocks. *Science* **291**, 851–853 (2001).
- Wang, J. F., Gudiksen, M. S., Duan, X. F., Cui, Y. & Lieber, C. M. Highly polarized photoluminescence and photodetection from single indium phosphide nanowires. *Science* **293**, 1455–1457 (2001).
- Cui, Y., Wei, Q. Q., Park, H. K. & Lieber, C. M. Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species. *Science* **293**, 1289–1292 (2001).
- Duan, X. F., Huang, Y., Cui, Y., Wang, J. F. & Lieber, C. M. Indium phosphide nanowires as building blocks for nanoscale electronic and optoelectronic devices. *Nature* **409**, 66–69 (2001).
- Huang, Y. *et al.* Logic gates and computation from assembled nanowire building blocks. *Science* **294**, 1313–1317 (2001).
- Bachtold, A., Hadley, P., Nakanishi, T. & Dekker, C. Logic circuits with carbon nanotube transistors. *Science* **294**, 1317–1320 (2001).
- Derycke, V., Martel, R., Appenzeller, J. & Avouris, P. Carbon nanotube inter- and intramolecular logic gates. *Nano Lett.* **1**, 453–456 (2001).
- Huang, Y., Duan, X. F., Wei, Q. Q. & Lieber, C. M. Directed assembly of one-dimensional nanostructures into functional networks. *Science* **291**, 630–633 (2001).
- Zhou, C. W., Kong, J., Yenilmez, E. & Dai, H. J. Modulated chemical doping of individual carbon nanotubes. *Science* **290**, 1552–1555 (2000).
- Gudiksen, M. S. & Lieber, C. M. Diameter-selective synthesis of semiconductor nanowires. *J. Am. Chem. Soc.* **122**, 8801–8802 (2000).
- Cui, Y., Lauhon, L. J., Gudiksen, M. S., Wang, J. F. & Lieber, C. M. Diameter-controlled synthesis of single-crystal silicon nanowires. *Appl. Phys. Lett.* **78**, 2214–2216 (2001).
- Gudiksen, M. S., Wang, J. F. & Lieber, C. M. Synthetic control of the diameter and length of single crystal semiconductor nanowires. *J. Phys. Chem. B* **105**, 4062–4064 (2001).
- Wagner, R. S. in *Whisker Technology* 47–119 (Wiley-Interscience, New York, 1970).
- Morales, A. M. & Lieber, C. M. A laser ablation method for the synthesis of crystalline semiconductor nanowires. *Science* **279**, 208–211 (1998).
- Duan, X. F. & Lieber, C. M. General synthesis of compound semiconductor nanowires. *Adv. Mater.* **12**, 298–302 (2000).
- Cui, Y., Duan, X. F., Hu, J. T. & Lieber, C. M. Doping and electrical transport in silicon nanowires. *J. Phys. Chem. B* **104**, 5213–5216 (2000).
- Nicewarner-Pena, S. R. *et al.* Submicrometer metallic barcodes. *Science* **294**, 137–141 (2001).
- Chow, E. *et al.* Three-dimensional control of light in a two-dimensional photonic crystal slab. *Nature* **407**, 983–986 (2000).
- Huang, M. H. *et al.* Room-temperature ultraviolet nanowire nanolasers. *Science* **292**, 1897–1899 (2001).
- Bachtold, A. *et al.* Scanned probe microscopy of electronic transport in carbon nanotubes. *Phys. Rev. Lett.* **84**, 6082–6085 (2000).
- Cui, Y., Ouyang, M., Yang, P. & Lieber, C. M. Controlled growth and electrical properties of heterojunctions of carbon nanotubes and silicon nanowires. *Nature* **399**, 48–51 (1999).
- Bennett, C. H. & DiVincenzo, D. P. Quantum information and computation. *Nature* **404**, 247–255 (2000).

Acknowledgements

We thank X. Duan for discussions, and W. MoberlyChan and A. J. Garratt-Reed for assistance with TEM imaging and analysis. M.S.G. thanks the NSF for predoctoral fellowship support. C.M.L. acknowledges support of this work by the Office of Naval Research and Defense Advanced Projects Research Agency.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.M.L. (e-mail: cml@cmliris.harvard.edu).

Ultralow-threshold Raman laser using a spherical dielectric microcavity

S. M. Spillane, T. J. Kippenberg & K. J. Vahala

Department of Applied Physics, California Institute of Technology, Pasadena, California 91125, USA

The ability to confine and store optical energy in small volumes has implications in fields ranging from cavity quantum electrodynamics to photonics. Of all cavity geometries, micrometre-sized dielectric spherical resonators are the best in terms of their ability to store energy for long periods of time within small volumes¹. In the sphere, light orbits near the surface, where long confinement times (high Q) effectively wrap a large interaction distance into a tiny volume. This characteristic makes such resonators uniquely suited for studies of nonlinear coupling of light with matter. Early work^{2,3} recognized these attributes through Raman excitation in microdroplets—but microdroplets have not been used in practical applications. Here we demonstrate a micrometre-scale, nonlinear Raman source that has a highly efficient pump–signal conversion (higher than 35%) and pump thresholds nearly 1,000 times lower than shown before. This represents a route to compact, ultralow-threshold sources for numerous wavelength bands that are usually difficult to access. Equally important, this system can provide a compact and simple building block for studying nonlinear optical effects and the quantum aspects of light.

Raman sources based on silica may be helpful in extending the available range of semiconductor lasers. But to date these sources have required very high pump power, and are macroscale devices. Silica microspheres are good candidates for next-generation Raman sources, and have the highest Q -factors (quality factors) of any optical resonator ($>10^9$). Here we report a Raman source consisting of a high- Q silica microsphere coupled to an optical fibre. This device enables a large reduction in the necessary threshold pump power⁴, while fibre-coupling⁵ notably improves overall efficiency and provides a convenient method of optical field transport. Although a single Raman laser is investigated in this work, the ability to fibre-couple should enable easy scaling to multiple resonant systems along a single fibre (Fig. 1).

The lasing threshold occurs when cavity round-trip gain equals round-trip loss. For an intensity-dependent gain coefficient (such as for a Raman laser), and taking into account the power build-up factor in a resonator, we obtain the following equation for threshold pump power;

$$P_{\text{threshold}} = \frac{\pi^2 n^2 V_{\text{eff}}}{\lambda_p \lambda_R \Gamma B g} Q_c^p \left(\frac{1}{Q_t^p} \right)^2 \frac{1}{Q_t^R} \quad (1)$$

Here $P_{\text{threshold}}$ denotes the incident power necessary in the fibre (not the power coupled into the resonator), n is the index of refraction, V_{eff} is the effective pump mode volume (that is, taking the full-width at half-maximum of the intensity distribution), λ_p and λ_R are the pump and Raman wavelengths, Γ is the spatial mode overlap factor between pump and Raman modes ($\Gamma \approx 1$), g is the nonlinear bulk Raman gain coefficient, B is a correction factor of the circulating power due to internal backscattering (between 1 and 0.5), and Q_t^p is the total quality factor for the pump mode, made up of an intrinsic contribution Q_0^p and a coupling contribution Q_c^p (and similarly for the Raman mode). Equation (1) has the important feature that the threshold pump power scales inversely with the factor Q^2/V , which is the same as quality factor multiplied by the cavity Purcell factor ($\propto Q/V$). Thus quality factor plays a dominant role in device performance, resulting from the necessity of a doubly resonant process. This is a classical conclusion, neglecting the

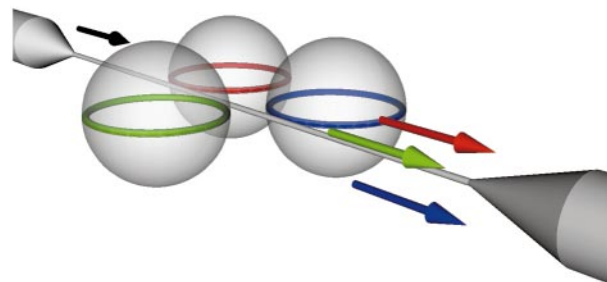


Figure 1 Cascading multiple Raman lasers along a single fibre. Black arrow represents input wave. Each microsphere may emit a different Raman wave, indicated by the coloured arrows.

additional benefit due to enhancement of the gain coefficient by cavity quantum electrodynamics⁶ (QED).

Silica microspheres were fabricated by melting the tip of a standard telecommunication (SMF-28) fibre with a CO₂ laser; surface tension creates a spherical volume with low eccentricities ($\sim 2\%$). These spheres were optically coupled by the use of a tapered optical fibre, formed by heating and stretching a length of single-mode fibre. Typical taper diameters were approximately 1.5 μm with losses of less than 5%. Details of the fabrication of microspheres and tapers are given elsewhere⁵. A polarization controller was used to adjust the incident polarization state at the sphere–taper junction to maximize the coupling efficiencies while preserving large values of Q . In these measurements, values of Q were in the low 10^8 range, believed to be limited by surface scattering and OH absorption⁷. Surface scattering also induces backscattering of power and couples the initially degenerate clockwise and anti-clockwise circulating modes, causing a splitting of the resonance wavelength⁸, in addition to a reduction of the circulating power (up to a factor of two). The taper position is controlled relative to the microsphere by a three-axis stage with a resolution of 20 nm. A tunable 1,550-nm external-cavity diode laser with 300-kHz linewidth is used as a pump. The laser is scanned repeatedly through a frequency range of approximately 60 GHz around a single whispering-gallery mode (WGM).

Figure 2 shows the emission spectrum for a Raman microsphere laser (intrinsic pump quality factor of $Q_0 = 10^8$) excited far above the threshold for stimulated Raman scattering. There are a multi-

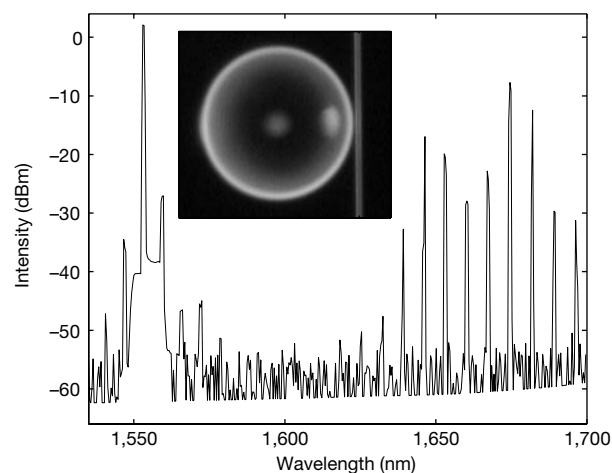


Figure 2 Spectrum of a 70- μm -diameter Raman microsphere laser with pump powers of 2 mW. The pump is at 1,555 nm. Peaks located around 1,670 nm are Raman oscillations, separated by the free spectral range of the microsphere. Secondary lines around 1,555 nm are due to four-wave mixing between the pump and two Raman waves. Inset, a microsphere coupled to a fibre taper.

tude of nonlinearly generated wavelengths, from stimulated Raman peaks centred around 1,670 nm to Raman-assisted four-wave-mixing peaks located symmetrically about the 1,550-nm pump. Stimulated Raman scattering has a very high threshold (nonlinear gain coefficient of 10^{-11} cm W $^{-1}$), requiring significant circulating pump powers. This suggests that other nonlinear processes having lower thresholds, such as stimulated Brillouin scattering (SBS), could also be present in the systems tested. SBS, for example, should have a threshold roughly 500 times lower than stimulated Raman scattering (nonlinear gain coefficient is 500 times larger). To determine the presence of SBS, an optical spectrum analyser was used to measure backward-propagating optical power coupled from the microresonator into the fibre taper. The gain spectrum for SBS is very narrow (<100 MHz) with a roughly 10 GHz frequency down-shift in silica. The microresonators tested here have free spectral ranges on the order of terahertz, with an eccentricity splitting of the azimuthal modes on the gigahertz scale⁹. In general, this splitting is dependent upon fabrication-induced irregularities, thus overlap with the SBS spectrum is unlikely. Consequently, SBS was only observed when a WGM overlapped the Brillouin gain spectrum, in agreement with experiments on liquid droplets¹⁰. However, the backward spectral monitor did show the expected strong Raman oscillation. Cascaded Raman peaks, if present, were not observable due to the wavelength range of our optical spectrum analyser. We note that four-wave mixing associated with the Kerr effect was also observed. However, such processes in microcavities are governed by strict conditions imposed by the combined effect of the phase-matching requirement with the WGM spectral structure. This will be investigated in future work.

To investigate the dependence on quality factor, the threshold was measured while varying the coupling between the taper and microsphere by changing the air gap (Fig. 3). The data follow a near-parabolic shape, with a measured minimum value of 62 μ W. This value is, to our knowledge, the lowest directly measured (not inferred) threshold for any nonlinear substance to date. A theoretical fit, based on exponential dependence of the coupling Q , shows excellent agreement. We also obtain good agreement with a calculated minimum threshold value of 50 μ W under the assumption that Raman mode Q is equal to pump mode Q . The minimum threshold does not lie at the critical coupling point¹¹, where

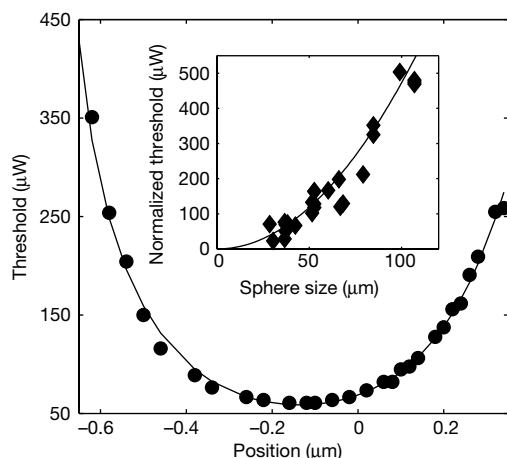


Figure 3 Coupling gap and size dependence of the Raman threshold. Main figure, Raman oscillation threshold versus taper-sphere gap for a 40- μ m-diameter sphere ($Q_0 = 10^6$). Position is measured from the critical coupling point, and negative values correspond to the undercoupled regime. The minimum threshold occurs with the microsphere about 0.15 μ m undercoupled, and corresponds to a transmission of 12%. Solid line, a theoretical fit to the threshold equation. Inset, normalized minimum threshold versus sphere size, which follows a quadratic dependence.

circulating pump power (Raman gain) is largest. It in fact occurs for the system in a slightly undercoupled state (weaker coupling than at the critical point), corresponding to an observed pump transmission of 12%. This shift results from the interplay between Raman gain and loss, caused by the coupling of both the pump and Raman wave to the fibre waveguide. In the simplified case where the quality factors of pump and oscillating modes are equal, the theoretical minimum occurs for a pump power transmission of 11%, in good agreement with the experimental value.

Examining the threshold equation (equation (1)), the threshold is expected to scale approximately as radius squared (effective mode volume is nearly quadratic in radius). To investigate this, the minimum threshold for various sphere diameters ranging from 28 to 110 μ m was measured (Fig. 3 inset). The threshold indeed follows a quadratic dependence on size. The dependence on quality factor is normalized out (assuming pump and Raman Q are identical) due to its strong effect on threshold. In previous work on microdroplets, it was found that for droplets with diameters of less than 30 μ m there is an additional threshold reduction due to cavity QED effects¹². In the present work we could not accurately determine threshold in this size range. Smaller spheres experienced very large temperature effects, causing the WGMs to experience instabilities¹³ as well as thermal shifting of the spectrum, complicating pumping of the resonant mode. With sufficient thermal control it may be possible to quantify precisely the reduction of threshold in this smaller size regime.

The efficiency of the Raman oscillator was investigated by decreasing the pump power until a single emission wavelength was observed on the optical spectrum analyser (Fig. 4). The measured threshold is 86 μ W for this 40- μ m sphere. This value is nearly 1,000 times lower than the corresponding values measured in previous work using microdroplets (45 mW for a 30- μ m CS₂ droplet; ref. 11), despite the fact that silica has a 1,000 times smaller Raman gain coefficient. This improvement of nearly 10^6 results from efficient single WGM excitation, whereas free-space beams (used previously) lack high spatial mode selectivity. Upon closer examination, the Raman emission in fact consisted of a number of separate simultaneously oscillating peaks (to the temporal resolution of the optical spectrum analyser), corresponding to different azimuthal modes. The number of oscillating modes was typically 3–5, owing to the spatial selectivity of the fibre-taper. The measured unidirectional absolute conversion efficiency was 16%, with a differential quantum efficiency (clockwise and anticlockwise

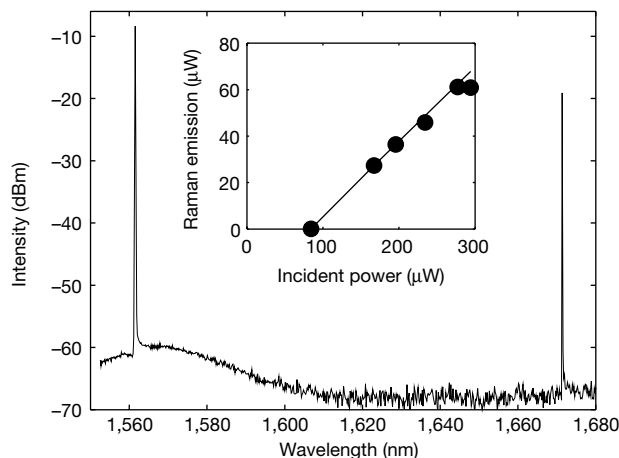


Figure 4 Single longitudinal mode Raman lasing. Raman spectrum for a 40- μ m-diameter microsphere, exhibiting a unidirectional conversion efficiency of 16% (pump is at 1,555 nm). Inset, Raman power output (sum of forward and backward emission) versus incident pump power. Differential quantum efficiency is 36%.

oscillation) of 36%. In other microspheres tested, Raman output powers as large as 200 μ W were obtained.

There are several ways to improve the performance of this system. Quality factors of 10^9 have been achieved previously in fused-silica microspheres¹⁴, which should lead to submicrowatt Raman thresholds. The use of small spheres (provided temperature issues can be resolved) should improve Raman conversion efficiencies under increased pump power due to the increase in free spectral range, which decreases the efficiency of secondary Raman lines and Raman-assisted four-wave mixing. Disks should allow true single-mode emission. Concerning the outlook for future work and the implications of our work beyond an ultra-efficient and compact Raman source: there are several potential avenues that could be useful in fundamental studies. First, we note that opportunities exist to explore cavity QED effects in smaller microsphere samples⁶. As noted above, this will require an improvement in the stabilization of the sphere temperature. Additionally, the ability to achieve ultra-low-loss coupling between the resonator and the fibre taper suggests that compound resonant systems could be studied by attachment of multiple resonators along a single fibre taper. As noted earlier, we have observed Raman and SBS oscillation as well as four-wave mixing in these systems. These nonlinearities can be accessed in a compact volume through a nearly ideal field transport channel (optical fibre) and field coupling junction (the taper). As such, this system can be viewed more broadly as a building block for study of a host of nonlinearities within a high-Q silica resonator, and potentially for generation of non-classical photon states^{15,16} and their efficient transport. Indeed, we believe the compact nature of the system combined with the power of fibre field transport could afford relatively straightforward access to single or tandem resonant systems in normally challenging environments such as ultralow-temperature chambers¹⁷. The ability to load or suitably modify spheres using dopants or quantum dots¹⁸ could also be useful in such studies. □

Received 10 October; accepted 7 December 2001.

1. Collot, L., Lefevre-Seguin, V., Brune, M., Raimond, J. M. & Haroche, S. Very high-Q whispering-gallery mode resonances observed on fused silica microspheres. *Europhys. Lett.* **23**, 327–334 (1993).
2. Qian, S. X. & Chang, R. K. Multiorder Stokes emission from micrometer-size droplets. *Phys. Rev. Lett.* **56**, 926–929 (1986).
3. Lin, H. B., Huston, A. L., Eversole, J. D. & Campillo, A. J. Double-resonance stimulated Raman-scattering in micrometer-sized droplets. *J. Opt. Soc. Am. B* **7**, 2079–2089 (1990).
4. Braunstein, D., Khazanov, A. M., Koganov, G. A. & Shuker, R. Lowering of threshold conditions for nonlinear effects in a microsphere. *Phys. Rev. A* **53**, 3565–3572 (1996).
5. Knight, J. C., Cheung, G., Jacques, F. & Birks, T. A. Phase-matched excitation of whispering-gallery-mode resonances by a fiber taper. *Opt. Lett.* **22**, 1129–1131 (1997).
6. Chang, R. K. & Campillo, A. J. (eds) *Optical Processes in Microcavities* (World Scientific, Singapore, 1996).
7. Gorodetsky, M. L., Savchenkov, A. A. & Ilchenko, V. S. Ultimate Q of optical microsphere resonators. *Opt. Lett.* **21**, 453–455 (1996).
8. Weiss, D. S. *et al.* Splitting of high-Q Mie modes induced by light backscattering in silica microspheres. *Opt. Lett.* **20**, 1835–1837 (1995).
9. Lai, H. M., Leung, P. T., Young, K., Barber, P. W. & Hill, S. C. Time-independent perturbation for leaking electromagnetic modes in open systems with application to resonances in microdroplets. *Phys. Rev. A* **41**, 5187–5198 (1990).
10. Zhang, J. Z. & Chang, R. K. Generation and suppression of stimulated Brillouin scattering in single liquid droplets. *J. Opt. Soc. Am. B* **6**, 151–153 (1989).
11. Cai, M., Painter, O. & Vahala, K. J. Observation of critical coupling in a fiber taper to a silica-microsphere whispering-gallery mode system. *Phys. Rev. Lett.* **85**, 74–77 (2000).
12. Lin, H. B. & Campillo, A. J. CW nonlinear optics in droplet microcavities displaying enhanced gain. *Phys. Rev. Lett.* **73**, 2440–2443 (1994).
13. Ilchenko, V. S. & Gorodetskii, M. L. Thermal nonlinear effects in optical whispering gallery microresonators. *Laser Phys.* **2**, 1004–1009 (1992).
14. Vernooij, D. W., Ilchenko, V. S., Mabuchi, H., Steed, E. W. & Kimble, H. J. High-Q measurements of fused-silica microspheres in the near infrared. *Opt. Lett.* **23**, 247–249 (1998).
15. Bachor, H.-A., Levenson, M. D., Walls, D. F., Perlmutter, S. H. & Shelby, R. M. Quantum nondemolition measurements in an optical-fiber ring resonator. *Phys. Rev. A* **38**, 180–190 (1988).
16. Silberhorn, Ch. *et al.* Generation of continuous variable Einstein-Podolsky-Rosen entanglement via the Kerr nonlinearity in an optical fiber. *Phys. Rev. Lett.* **6**, 4267–4270 (2001).
17. Treussart, F. *et al.* Evidence of the intrinsic Kerr bistability of high-Q microsphere resonators in superfluid helium. *Eur. Phys. J. D* **1**, 235–238 (1998).
18. Fan, X., Palinginis, P., Lacey, S., Wang, H. & Lonergan, M. C. Coupling semiconductor nanocrystals to a fused-silica microsphere: a quantum-dot microcavity with extremely high Q factors. *Opt. Lett.* **25**, 1600–1602 (2000).

Acknowledgements

We thank A. D. Stone and R. K. Chang for comments. This work was supported by DARPA, NSF and the Caltech Lee Center.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.J.V. (email: vahala@caltech.edu).

Observation of ligand effects during alkene hydrogenation catalysed by supported metal clusters

A. M. Argo, J. F. Odzak, F. S. Lai & B. C. Gates

Department of Chemical Engineering and Materials Science, University of California, Davis, California 95616, USA

Homogeneous organometallic catalysts and many enzymes activate reactants through coordination to metal atoms; that is, the reactants are turned into ligands and their reactivity controlled through other ligands in the metal's coordination sphere¹. In the case of supported metal clusters, catalytic performance is influenced by the support and by adsorbed reactants, intermediates or products. The adsorbates are usually treated as ligands, whereas the influence of the supports is usually ascribed to electronic interactions^{2,3}, even though metal clusters supported on oxides^{4–6} and zeolites⁷ form chemical bonds to support oxygen atoms. Here we report direct observations of the structure of supported metal clusters consisting of four iridium atoms, and the identification of hydrocarbon ligands bound to them during propene hydrogenation. We find that propene and molecular hydrogen form propylidyne and hydride ligands, respectively⁸, whereas simultaneous exposure of the reactants to the supported iridium cluster yields ligands that are reactive intermediates during the catalytic propane-formation reaction. These intermediates weaken the bonding within the tetrahedral iridium cluster and the interactions between the cluster and the support, while replacement of the MgO support with γ -Al₂O₃ boosts the catalytic activity tenfold, by affecting the bonding between the reactant-derived ligands and the cluster and therefore also the abundance of individual ligands. This interplay between the support and the reactant-derived ligands, whereby each influences the interaction of the metal cluster with the other, shows that the catalytic properties of supported metal catalysts can be tuned by careful choice of their supports.

Our Ir₄ clusters resemble the platinum nanoclusters in industrial naphtha reforming catalysts^{9,10}. Decarbonylation of [Ir₄(CO)₁₂] on porous γ -Al₂O₃ (surface area, 100 m² g⁻¹) or of [HIr₄(CO)₁₁]⁻ on MgO (70 m² g⁻¹)^{11,12} yielded supported Ir₄ catalyst samples with 1 wt% Ir that were used to catalyse the hydrogenation of ethene and propene. Conventional measurements of catalytic reaction rates¹³ were complemented by extended X-ray absorption fine structure (EXAFS)¹⁴ and infrared (IR)⁸ spectroscopies, allowing us to determine the cluster structure and identify the ligands derived from the supports and the reactants while the catalysts were active in flow reactors. Infrared spectra of adsorbate ligands were determined by subtracting the spectrum of the sample in an inert gas (He) from that of the sample in the treatment gas; when absorption by the reactant gas or product gas was significant, its spectrum was also subtracted⁸.

We find that the catalytic activity of $\text{Ir}_4/\gamma\text{-Al}_2\text{O}_3$ is significantly greater than that of Ir_4/MgO , but less than that of the extended metal surface $\text{Pt}(111)$ (refs 15 and 16; Table 1).

During treatment in either H_2 or alkene^{8,13} and during catalytic hydrogenation of ethene or propene at 240–298 K, the EXAFS first-shell Ir–Ir coordination numbers remain close to 3 (Table 2), indicating that the Ir_4 clusters on the different supports remain tetrahedral. The absence of higher Ir–Ir shell peaks in the spectra is consistent with the clusters being isolated. Contributions to EXAFS spectra of Ir–O at a distance of $2.1 \pm 0.1 \text{ \AA}$ are consistent with covalent bonding between the cluster and the support^{4–7}, while longer Ir–O contributions at $2.6 \pm 0.1 \text{ \AA}$ (Table 2) suggest non-bonding interactions of Ir_4 , perhaps with OH groups on the support surface. The Ir–O coordination numbers (Table 2) suggest the supports both act as polydentate ligands⁷, but differing in the number of bonds they form between the cluster and surface oxygen atoms.

The IR spectra show that the nature of the alkene-derived ligands coordinating to Ir_4 in the absence of H_2 varies with support: propene reacts with $\gamma\text{-Al}_2\text{O}_3$ -supported Ir_4 at 298 K to give $\text{CH}_3\text{CH}_2\text{C}(\text{Ir}_4)/\gamma\text{-Al}_2\text{O}_3$ (cluster-bound propylidyne)⁸ (Supplementary Information Fig. 1), but is only physisorbed on MgO -supported Ir_4 . The H_2 adsorption kinetics observed when exposing the catalysts to H_2 on its own is consistent with iridium hydride formation, but the IR and ^1H NMR spectra are not sensitive enough to directly identify the hydride. (Hydride formation is observed with $\text{Rh}_6/\text{La}_2\text{O}_3$; A.M.A. and B.C.G., unpublished results).

The ligands on Ir_4 observed by IR spectroscopy during steady-state alkene hydrogenation include hydrocarbons not observed with the alkene alone (Supplementary Information Table 1), with the nature of the most abundant ligands varying with support and reaction conditions. Individual ligands were identified, and spectra of combinations of ligands resolved, by varying the reaction conditions and comparing against published spectra^{17–20} (Supplementary Information Figs 1 and 2 and Table 1). For example, ethyl, π -bonded ethene, ethylidyne, and di- σ -bonded ethene formed on $\text{Ir}_4/\gamma\text{-Al}_2\text{O}_3$ during ethene hydrogenation at 288 K and 760 torr (ethene partial pressure ($p_{\text{C}_2\text{H}_4}$), 206 torr; p_{H_2} , 139–397 torr; the balance He). With changing reaction conditions, the band intensities of some ligands changed whereas others remained nearly constant (Fig. 1).

A ligand is a reaction intermediate if its concentration can be correlated with the catalytic activity of the system. For example, as p_{H_2} was increased, the intensities of the IR bands representing π -bonded ethene, ethylidyne, and di- σ -bonded ethene on $\text{Ir}_4/\gamma\text{-Al}_2\text{O}_3$ remained nearly constant, whereas the intensities of the ethyl bands increased almost proportionally with p_{H_2} (Fig. 1) and catalytic activity (Fig. 2). A similar trend was observed for ethyl on Ir_4/MgO (Supplementary Information Fig. 3). Thus, ethyl is identified as a reaction intermediate under these conditions, whereas π -

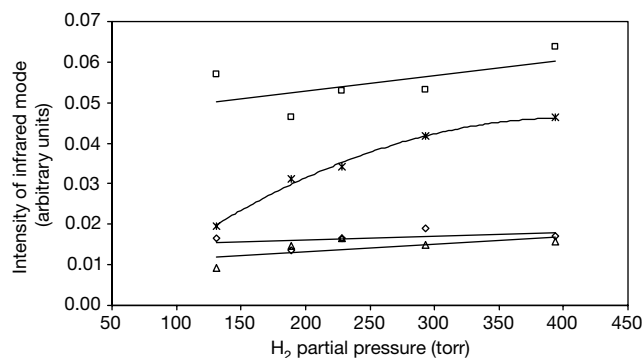


Figure 1 Infrared spectra of adsorbate ligands on $\text{Ir}_4/\gamma\text{-Al}_2\text{O}_3$ during ethene hydrogenation. Reaction conditions were 288 K and a gas pressure of 760 torr ($p_{\text{C}_2\text{H}_4}$, 206 torr; p_{H_2} , 131–393 torr; balance He). The infrared intensities corresponding to various ligands were obtained from the following bands: π -bonded ethene, $3,055 \text{ cm}^{-1}$ (diamond data symbols); di- σ -bonded ethene, $2,988 \text{ cm}^{-1}$ (squares); ethyl, $2,921$ and $2,870 \text{ cm}^{-1}$ (asterisks); and ethylidyne, $2,890 \text{ cm}^{-1}$ (triangles). The adsorbates are inferred to have been bonded to the clusters and not significantly to the supports, because of the similarities in peak positions to those of equivalent ligands on molecular metal clusters and those of equivalent adsorbates on single crystals of metal, and because the supports themselves adsorbed negligible amounts of hydrogen or alkene relative to samples consisting of the supported metal clusters.

bonded ethene, ethylidyne, and di- σ -bonded ethene are either ‘spectators’ or involved in virtually equilibrated elementary steps in the catalytic cycle.

We observed that at $p_{\text{C}_2\text{H}_4} \leq 200$ torr, increasing $p_{\text{C}_2\text{H}_4}$ gave lower catalytic reaction rates, whereas the reverse was true at higher $p_{\text{C}_2\text{H}_4}$. Infrared spectra showed that at $p_{\text{C}_2\text{H}_4} \geq 200$ torr, a new adsorbate, π -bonded ethene, was present, and as $p_{\text{C}_2\text{H}_4}$ increased, its band intensity increased. Thus, π -bonded ethene is implicated as a reactive intermediate at the higher $p_{\text{C}_2\text{H}_4}$. The competition between reaction intermediates on the cluster is similar to that on $\text{Pd}(111)$ (ref. 21), except that the surface intermediates were π -bonded ethene and di- σ -bonded ethene.

In propene hydrogenation at $p_{\text{H}_2} = 300$ torr, variation of $p_{\text{C}_3\text{H}_6}$ caused the 1-propyl band intensities to vary, along with the catalytic reaction rate; the intensity of the π -bonded propene band also correlates with the rate (Supplementary Information Figs 4 and 5), indicating that it too is a significant reaction intermediate. The similarity of the results obtained for ethene hydrogenation and propene hydrogenation implies that the two reactions proceed by similar mechanisms on supported Ir_4 (for example, Fig. 3 for propene on Ir_4/MgO), as they do on metal surfaces^{15,16}.

The data also identify some differences between ethene hydrogenation and propene hydrogenation: on supported Ir_4 , propy-

Table 1 Activities of catalysts for alkene hydrogenation

Catalyst	Reactant alkene	Catalyst pretreatment	Reference	TOF (s^{-1})	Reaction conditions*
$\text{Pt}(111)$	Propene	None	15	8.6	a
$\text{Pt}(111)$	Propene	Propene	15	8.4	a
$\text{Ir}_4/\gamma\text{-Al}_2\text{O}_3$	Propene	None	This work	0.21	a
Ir_4/MgO	Propene	None	This work	0.091	a
$\text{Ir}_{\text{agg}}/\gamma\text{-Al}_2\text{O}_3$	Propene	None	This work	0.37	a
$\text{Ir}_4/\gamma\text{-Al}_2\text{O}_3$	Propene	None	This work	0.11	b
$\text{Ir}_4/\gamma\text{-Al}_2\text{O}_3$	Propene	Propene	This work	0.007	b
Ir_4/MgO	Propene	None	This work	0.011	b
Ir_4/MgO	Propene	Propene	This work	0.009	b
$\text{Pt}(111)$	Ethene	None	16	12	a
$\text{Ir}_4/\gamma\text{-Al}_2\text{O}_3$	Ethene	None	This work	0.23	a
Ir_4/MgO	Ethene	None	This work	0.052	a
$\text{Ir}_{\text{agg}}/\gamma\text{-Al}_2\text{O}_3$	Ethene	None	This work	0.34	a

Catalytic activities are reported as rate of reaction per exposed metal atom (turnover frequency, TOF); for Ir_4 , these are rates per total Ir atom. The rates for $\text{Pt}(111)$ are reported per surface Pt atom. Ir_{agg} refers to non-uniform aggregates of Ir with an average diameter of $\sim 13 \text{ \AA}$, formed by treatment of $[\text{Ir}_4(\text{CO})_{12}]/\gamma\text{-Al}_2\text{O}_3$ in H_2 at 673 K for 2 h.

* a, 295 K ($p_{\text{H}_2} = 100$ torr, $p_{\text{C}_2\text{H}_4} = 40$ torr, $p_{\text{H}_2\text{O}} = 617$ torr); b, 298 K ($p_{\text{H}_2} = 18$ torr, $p_{\text{C}_2\text{H}_4} = 6.8$ torr, $p_{\text{H}_2\text{O}} = 755$ torr).

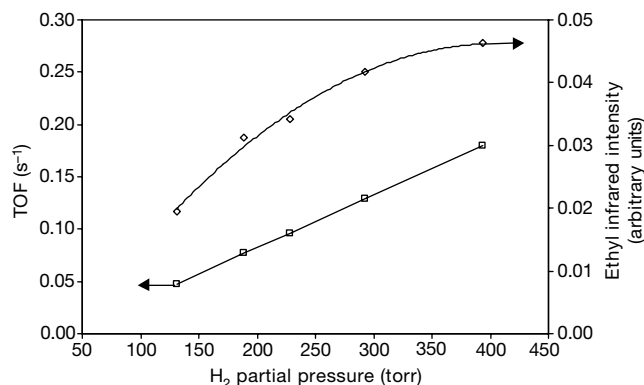


Figure 2 Catalytic activity and average ethyl band intensity as a function of p_{H_2} during ethene hydrogenation catalysed by $Ir_4/\gamma-Al_2O_3$. The data show the dependence of catalytic activity (turnover frequency, TOF) and average ethyl band intensity (determined from the bands at 2,921 and 2,870 cm^{-1}) on p_{H_2} during catalysis at 288 K and 760 torr (p_{H_2} , 131–394 torr; $p_{C_2H_4}$, 206 $\pm$ 5 torr; balance He). The correlation between the band intensity and the catalytic activity indicates that ethyl is a reactive intermediate in the catalytic reaction.

lidyne and 1-propyl are predominant during catalysis at $p_{C_3H_6} < 75$ torr, whereas π -bonded propene is observed (with propylidyne and 1-propyl) at higher $p_{C_3H_6}$. The concentrations of π -bonded propene, 1-propyl, and propylidyne on the clusters during catalysis were approximately independent of p_{H_2} in the range 50–400 torr. The clusters became essentially saturated with propene-derived ligands under these conditions. In contrast, the clusters were only partially covered with ethene-derived ligands during ethene hydrogenation, as shown by a strong dependence of the IR band intensities on p_{H_2} .

EXAFS spectra complement IR spectra by showing how changes in the reactant-derived ligands influence the cluster frame and its interaction with the support. Contacting of $Ir_4/\gamma-Al_2O_3$ with alkene or H_2 alone led to no significant changes in the Ir–Ir or Ir–O_{support} distances. However, during catalysis, the Ir–Ir distance and the longer, non-bonding Ir–O_{support} distance exceed the corresponding distances observed under non-catalytic conditions. (The shorter, bonding Ir–O_{support} distance does not show a significant trend.) Changes in p_{H_2} led to changes in the Ir–Ir and the longer Ir–O_{support} distances in $Ir_4/\gamma-Al_2O_3$, and these distances increased with increasing catalytic activity (Table 2). Furthermore, the Ir–Ir distance increased with the IR band intensities of ethyl on Ir_4 (Supplementary Information Fig. 6), suggesting that ethyl ligands cause the

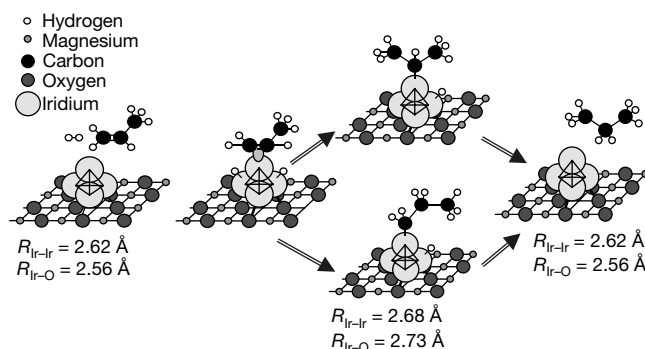


Figure 3 Schematic representation of the hydrogenation of propene on MgO-supported Ir_4 . Propene is initially π -bonded to the cluster then hydrogenated to give 1-propyl or 2-propyl, which is hydrogenated to give propane. The Ir–Ir and longer (non-bonding) Ir–O distances change as shown.

changes in the cluster structure and cluster/support interaction. (However, hydride ligands, also identified as reactive intermediates, and OH groups that form on the support from the hydride ligands might also contribute to these changes.) Increasing the reaction temperature from 256 to 285 K while keeping the flow rates of H_2 and propene constant increased the catalytic activity of Ir_4/MgO , the Ir–Ir distance (from 2.63 to 2.68 Å) and the non-bonding Ir–O_{support} distance (from 2.56 to 2.73 Å), implying changes in catalyst structure induced by a reaction intermediate (Table 2).

MgO and $\gamma-Al_2O_3$ differ in their electron-donor characteristics, and thus in their effects on the interaction of the clusters with reactant-derived ligands. The strongly basic MgO increases the charge density on Ir_4 more than the less basic $\gamma-Al_2O_3$ does, and consequently MgO reduces the strength of the alkene–cluster interaction—and thereby the catalytic activity—more than $\gamma-Al_2O_3$. Ir_4 on $\gamma-Al_2O_3$, for example, reacts with propene as extended metal surfaces do, yielding strongly bonded propylidyne^{17,20}, although Ir_4 on MgO does not. Weakening of the interactions between hydrocarbon ligands and Ir_4 caused by MgO is also reflected by the C–H stretching frequency seen in π -bonded ethene: the frequency is centred around 3,090 cm^{-1} for π -bonded ethene coordinated to Ir_4/MgO , and around 3,060 cm^{-1} for π -bonded ethene coordinated to $Ir_4/\gamma-Al_2O_3$. Similarly, the C–H stretching frequency of ethene di- σ -bonded to Ir_4/MgO is higher than that of the corresponding ligand attached to $Ir_4/\gamma-Al_2O_3$ (2,992–3,003 cm^{-1} versus 2,987 cm^{-1} , respectively).

The strong ligand effects exerted by the supports distinguish

Table 2 Structural parameters characterizing supported Ir_4 alkene hydrogenation catalysts

Support	Reactant alkene	p_{H_2} (mbar)	Conditions	T (K)	TOF (s ⁻¹)	N_{Ir-Ir}	R_{Ir-Ir} (Å)	$N_{Ir-O_s}^*$	R_{Ir-O_s} (Å)	$N_{Ir-O_l}^\dagger$	R_{Ir-O_l} (Å)
$\gamma-Al_2O_3$	None	0	He‡	298	—	3.2	2.62	0.8	2.01	1.1	2.61
$\gamma-Al_2O_3$	Ethene	175	Catalysis§	298	0.13	3.2	2.65	0.0	—	1.6	2.63
$\gamma-Al_2O_3$	Ethene	255	Catalysis	298	0.13	3.2	2.66	0.2	2.03	1.3	2.65
$\gamma-Al_2O_3$	Ethene	304	Catalysis	298	0.18	3.3	2.67	0.6	2.04	1.1	2.67
$\gamma-Al_2O_3$	Ethene	391	Catalysis	298	0.19	3.3	2.68	0.4	2.23	1.0	2.68
$\gamma-Al_2O_3$	None	0	He‡	261	—	3.2	2.64	0.6	2.08	0.7	2.61
$\gamma-Al_2O_3$	Propene	415	Catalysis	252	0.035	3.2	2.68	0.2	2.06	0.9	2.65
$\gamma-Al_2O_3$	Propene	415	Catalysis	259	0.068	3.3	2.69	0.5	2.04	0.9	2.66
$\gamma-Al_2O_3$	Propene	415	Catalysis	265	0.12	3.3	2.69	0.3	2.09	0.9	2.66
$\gamma-Al_2O_3$	Propene	415	Catalysis	281	0.23	3.3	2.69	0.3	2.11	1.0	2.68
MgO	None	0	He	242	—	2.9	2.62	1.8	2.07	0	—
MgO	Propene	358	Catalysis	256	0.034	2.7	2.63	1.6	2.06	1.1	2.56
MgO	Propene	358	Catalysis	264	0.043	2.7	2.64	1.6	2.08	1.1	2.66
MgO	Propene	358	Catalysis	270	0.084	2.8	2.65	1.4	2.09	0.8	2.70
MgO	Propene	358	Catalysis	285	0.20	3.1	2.68	1.4	2.09	0.9	2.73

Data collection and analysis as before⁸. T, temperature; N, EXAFS coordination number (error approximately $\pm$ 15%); R, EXAFS interatomic distance (error approximately $\pm$ 1%).

* Ir–O_s refers to the shorter Ir–O_{support} contribution.

† Ir–O_l refers to the longer Ir–O_{support} contribution.

‡ He flow of 12 ml min⁻¹ at 760 torr and 298 K.

§ Total gas flow of 30 ml min⁻¹ ($p_{C_2H_4}$ = 200 torr, the indicated p_{H_2} , and the balance He) at 760 torr and 298 K.

|| Total gas flow of 30 ml min⁻¹ ($p_{C_3H_6}$ = 135 torr, the indicated p_{H_2} , and the balance He) at 760 torr and 298 K.

supported metal clusters from extended metal surfaces. These effects (and thus opportunities for tuning catalytic properties by choice of the support) are most pronounced for the smallest clusters, and may be negligible for large supported metal particles for which only a small fraction of the metal atoms are bonded to the support.

Purely geometric effects can also distinguish the catalytic activity seen with small metal clusters from that seen with bulk metal or larger supported particles, by limiting the structures that can bond to a very small cluster and subsequently react on it. For example, propylidyne is stable on Ir₄/γ-Al₂O₃ when treated in He or H₂ at temperatures up to 523 K (ref. 8), whereas propylidyne on extended metal surfaces^{17,20} decomposes thermally under vacuum at 403–433 K and is largely hydrogenated in the presence of H₂ at room temperature. This qualitative difference in reactivity is attributed to the presence of neighbouring metal centres that facilitate reaction on the extended surfaces, and the lack of such centres on isolated Ir₄ (ref. 8). Thus, propylidyne on Pt(111) undergoes catalytic hydrogenation¹⁵ whereas propylidyne at saturation concentration on Ir₄/γ-Al₂O₃ does not react, rendering the clusters almost inactive for propene hydrogenation (Table 1). □

Received 16 August; accepted 11 December 2001.

- Collman, J. P., Hegedus, L. S., Norton, J. R. & Finke, R. G. *Principles and Applications of Organotransition Metal Chemistry* 2nd edn (University Science Books, Mill Valley, California, 1987).
- Stevenson, S. A., Dumesic, J. A., Baker, R. T. K. & Ruckenstein, E. (eds) *Metal-Support Interactions in Catalysis, Sintering, and Redispersion* (van Nostrand Reinhold, New York, 1987).
- Haller, G. L. & Resasco, D. E. Metal-support interaction: group VIII metals and reducible oxides. *Adv. Catal.* **36**, 173–235 (1989).
- Yudanov, I. V., Vent, S., Neyman, K., Pacchioni, G. & Rösch, N. Adsorption of Pd atoms and Pd-4 clusters on the MgO(001) surface: a density functional study. *Chem. Phys. Lett.* **275**, 245–252 (1997).
- Matveev, A. V., Neyman, K., Pacchioni, G. & Rösch, N. Density functional study of M-4 clusters (M = Cu, Ag, Ni, Pd) deposited on the regular MgO(001) surface. *Chem. Phys. Lett.* **299**, 603–612 (1999).
- Goellner, J. V. *et al.* Ligand-free osmium clusters supported on MgO: a density functional study. *Langmuir* **16**, 2736–2743 (2000).
- Ferrari, A. M. *et al.* Faujasite-supported Ir₄ clusters: A density functional model study of metal-zeolite interactions. *J. Phys. Chem. B* **103**, 5311–5319 (1999).
- Argo, A. M., Goellner, J. F., Phillips, B. L., Panjabi, G. A. & Gates, B. C. Reactivity of site-isolated metal clusters: propylidyne on γ-Al₂O₃-supported Ir₄. *J. Am. Chem. Soc.* **123**, 2275–2283 (2001).
- McVicker, G. B. *et al.* Effect of sulfur on the performance and on the particle size and location of platinum in Pt/KL hexane aromatization catalyst. *J. Catal.* **139**, 48–61 (1993).
- Jentoft, R. E., Tsapatsis, M., Davis, M. E. & Gates, B. C. Platinum clusters supported in zeolite LTL: influence of catalyst morphology on performance in *n*-hexane reforming. *J. Catal.* **179**, 565–580 (1998).
- Xu, Z. *et al.* Size-dependent catalytic activity of supported metal clusters. *Nature* **372**, 346–348 (1994).
- Gates, B. C. Supported metal clusters: synthesis, structure, and catalysis. *Chem. Rev.* **95**, 511–522 (1995).
- Argo, A. M. *Influence of Supports, Cluster Structure, and Cluster Composition on Hydrogenation Reactions Catalyzed by Oxide-Supported Metal Clusters*. Thesis, Univ. California at Davis (2001).
- Odzak, J. F., Argo, A. M., Lai, F. S. & Gates, B. C. A flow through X-ray absorption spectroscopy cell for characterization of powder catalysts in the working state. *Rev. Sci. Instrum.* **72**, 3943–3945 (2001).
- Cremer, P. S., Su, X., Shen, Y. R. & Somorjai, G. A. Hydrogenation and dehydrogenation of propylene on Pt(111) studied by sum frequency generation from UHV to atmospheric pressure. *J. Phys. Chem.* **100**, 16302–16309 (1996).
- Cremer, P. S., Su, X., Shen, Y. R. & Somorjai, G. A. Ethylene hydrogenation on Pt(111) monitored in situ at high pressure using sum frequency generation. *J. Am. Chem. Soc.* **118**, 2942–2949 (1996).
- Shahid, G. & Sheppard, N. Infrared spectra and the structures of the chemisorbed species resulting from the adsorption of propene and propane on a Pt/SiO₂ catalyst. *Spectrochim. Acta. A* **46**, 999–1010 (1990).
- Newell, H. E., McCoustra, M. R. S., Chesters, M. A. & De La Cruz, C. The thermal chemistry of adsorbed ethyl on the Pt(111) surface: infrared evidence for an ethylidene intermediate in the ethyl to ethylidene conversion. *J. Chem. Soc. Faraday Trans.* **94**, 3695–3698 (1998).
- Bent, B. E., Mate, C. M., Crowell, J. E., Koel, B. E. & Somorjai, G. A. Bonding and thermal decomposition of propylene, propadiene, and methylacetylene on the Rh(111) single crystal surface. *J. Phys. Chem.* **91**, 1493–1502 (1987).
- Chesters, M. A. *et al.* Infrared spectroscopic comparison of the chemisorbed species from ethene, propene, but-1-ene and *cis*- and *trans*-but-2-ene on Pt(111) and on a platinum/silica catalyst. *J. Chem. Soc. Faraday Trans.* **86**, 2757–2763 (1990).
- Neurock, M. & van Santen, R. A. A first principles analysis of C–H bond formation in ethylene hydrogenation. *J. Phys. Chem. B* **104**, 11127–11145 (2000).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank the US National Science Foundation for support and the National Synchrotron Light Source at Brookhaven National Laboratory for beam time.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.C.G. (email: bcgates@ucdavis.edu).

Towards robust regional estimates of CO₂ sources and sinks using atmospheric transport models

Kevin Robert Gurney*, Rachel M. Law†, A. Scott Denning*, Peter J. Rayner‡, David Baker‡, Philippe Bousquet§, Lori Bruhwiler||, Yu-Han Chen¶, Philippe Ciais§, Songmiao Fan#, Inez Y. Fung*, Manuel Gloor**, Martin Heimann*, Kaz Higuchi††, Jasmin John*, Takashi Maki‡‡, Shamil Maksyutov§§, Ken Masarie||, Philippe Peylin§, Michael Prather|||, Bernard C. Pak|||, James Randerson¶¶, Jorge Sarmiento#, Shoichi Taguchi##, Taro Takahashi*☆ & Chiu-Wai Yuen**

* Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado 80523, USA

† CSIRO Atmospheric Research, PMB 1, Aspendale, Victoria 3195, Australia

‡ National Center for Atmospheric Research (NCAR), Boulder, Colorado 80303, USA

§ Laboratoire des Sciences du Climat et de l'Environnement (LSCE), F-91198 Gif-sur-Yvette Cedex, France

|| Climate Monitoring and Diagnostics Laboratory, National Oceanic and Atmospheric Administration (NOAA), 326 Broadway R/CGI, Boulder, Colorado 80303, USA

¶ Department of Earth, Atmospheric, and Planetary Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02141, USA

AOS Program, Princeton University, Sayre Hall, Forrestal Campus, PO Box CN710, Princeton, New Jersey 08544-0710, USA

☆ Center for Atmospheric Sciences, McCone Hall, University of California, Berkeley, California 94720-4767, USA

** Max-Planck-Institut für Biogeochemie, D-07701 Jena, Germany

†† Meteorological Service of Canada, Environment Canada, Toronto, Ontario M3H 5T4, Canada

‡‡ Quality Assurance Section, Atmospheric Environment Division, Observations Department, Japan Meteorological Agency, 1-3-4 Otemachi, Chiyoda-ku, Tokyo 100-8122, Japan

§§ Institute for Global Change Research, Frontier Research System for Global Change, Yokohama 236-0001, Japan

||| Earth System Science, University of California, Irvine, California 92697-3100, USA

¶¶ Divisions of Engineering and Applied Science and Geological and Planetary Sciences, California Institute of Technology, Mail Stop 100-23, Pasadena, California 91125, USA

National Institute of Advanced Industrial Science and Technology, 16-1 Onogawa Tsukuba, Ibaraki 305-8569, Japan

*☆ Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York 10964, USA

Information about regional carbon sources and sinks can be derived from variations in observed atmospheric CO₂ concentrations via inverse modelling with atmospheric tracer transport models. A consensus has not yet been reached regarding the size and distribution of regional carbon fluxes obtained using this approach, partly owing to the use of several different atmospheric transport models^{1–9}. Here we report estimates of surface-atmosphere CO₂ fluxes from an intercomparison of atmospheric CO₂ inversion models (the TransCom 3 project), which includes 16 transport models and model variants. We find an uptake of CO₂

in the southern extratropical ocean less than that estimated from ocean measurements, a result that is not sensitive to transport models or methodological approaches. We also find a northern land carbon sink that is distributed relatively evenly among the continents of the Northern Hemisphere, but these results show some sensitivity to transport differences among models, especially in how they respond to seasonal terrestrial exchange of CO_2 . Overall, carbon fluxes integrated over latitudinal zones are strongly constrained by observations in the middle to high latitudes. Further significant constraints to our understanding of regional carbon fluxes will therefore require improvements in transport models and expansion of the CO_2 observation network within the tropics.

We estimate annual average fluxes for the 1992–96 period using each transport model and a common inversion set-up (see Methods). Methodological choices for this ‘control’ inversion have been selected on the basis of knowledge gained from a wide range of sensitivity tests (to be reported elsewhere). Performing the inversion with multiple transport models gives mean estimated fluxes that are relatively insensitive to reasonable variations in the set-up—and estimated uncertainties that represent a more complete estimate of the true uncertainty. The maximum number of regions in our inversion and the spatial distributions of fluxes within each region are fixed, precluding sensitivity tests of these inversion components.

Figure 1 shows the mean flux estimates (left-hand cross in each box) and two uncertainty measures for the control inversion. The first uncertainty measure is the mean of the individual model flux uncertainties (circles) which we designate the ‘within-model’ uncertainty. For any region, this estimated flux uncertainty must be smaller than the prior flux uncertainty (outer bounds of the boxes). The magnitude of the decrease indicates the degree to which the final flux estimate is constrained by the measurements. Figure 1 shows that the northern land regions and Australia are better constrained by the measurements than are the remaining land regions. The Southern Ocean region is well constrained by the atmospheric measurements, in part because it is treated as a single large region. The Atlantic regions are constrained more by their prior flux uncertainties, which are relatively small due to better coverage of ocean measurements in these regions.

The second uncertainty measure is the standard deviation of the flux estimates over the ensemble of models (error bars in Fig. 1). We call this the ‘between-model’ uncertainty. This measure indicates the degree to which transport model differences contribute to the range of flux estimates. Large between-model uncertainties are found for northern Africa, tropical America, temperate Asia and boreal Asia (all greater than 0.5 Gt C yr^{-1}).

For most regions, the between-model uncertainties are of similar or smaller magnitude than the within-model uncertainties. This suggests that the choice of transport model is not the critical determinant of the inferred fluxes. Comparing the uncertainties between regions indicates where the inversion would benefit most from new observations, and where model improvements are most needed. In this particular inversion, new measurements would be most useful over tropical continents and in the South America and South Atlantic regions, while the focus for resolving transport differences would be the northern and tropical land regions.

Regarding the model mean flux estimates, two results deserve attention. First, we find consistency between the ocean fluxes predicted in this study and those based on a global database¹⁰ of CO_2 partial pressure (p_{CO_2}), except in the Southern Ocean where the carbon uptake estimated here is roughly half that based on the p_{CO_2} database. This shift in uptake from south to north is required to match simultaneously large-scale concentration gradients (Fig. 2) and growth rates.

The mismatch between atmospheric and ocean estimates of the

Southern Ocean fluxes has been noted a decade ago¹¹. Our sensitivity tests find that the near-uniformity of observed concentration in the Southern Hemisphere and the small uncertainty associated with those measurements make this result robust to the choice of observing network, prior flux estimates, global ocean constraint, and transport (see Fig. 2 in Supplementary Information). The discrepancy also cannot be explained by a systematic bias in transport models, as the north–south transport has been investigated in a recent intercomparison¹² where successful simulations of the observed meridional gradient in SF_6 suggested reasonable veracity in gross interhemispheric transport.

One possible reconciliation between the p_{CO_2} database and the inverse result presented here is suggested by recent ocean measurements taken during January and August 2000 in the Indian

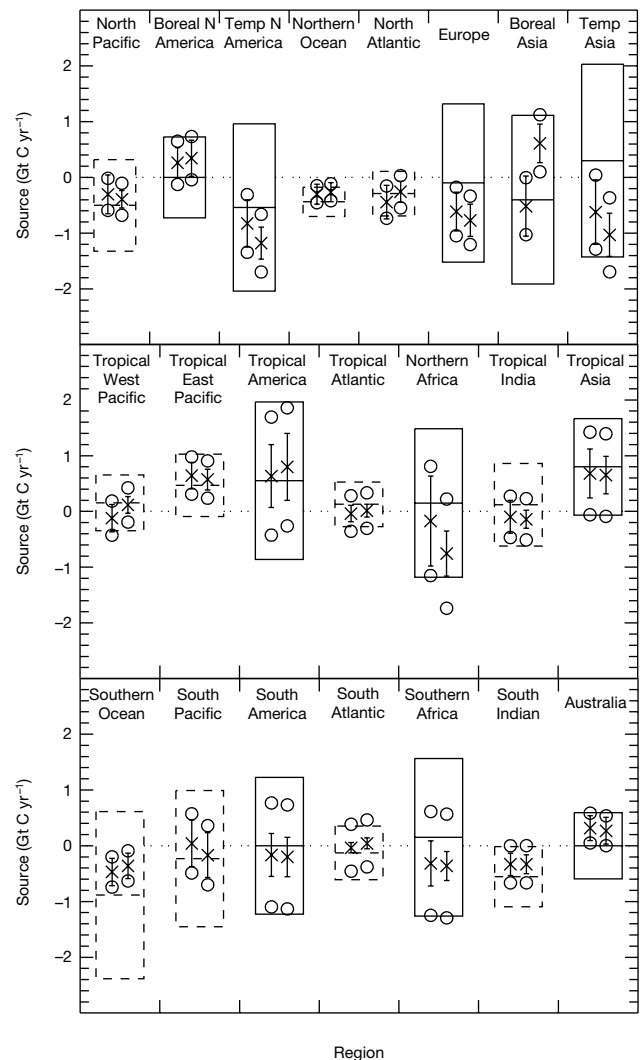


Figure 1 Mean estimated sources and uncertainties for two inversions. Left-hand symbols in each box are for the control inversion, right-hand symbols are for an inversion without the background seasonal biosphere flux. Mean estimated fluxes are shown by crosses, and include all background fluxes except fossil fuel. Positive values indicate a source to the atmosphere. The prior flux estimates and their uncertainties are indicated by the boxes (solid for land, dashed for ocean); the central horizontal bar indicates the prior flux estimate, and the top and bottom of the box give the prior flux uncertainty range. The mean estimated uncertainty across all models (the ‘within-model’ uncertainty) is indicated by the circles. The standard deviation of the models’ estimated fluxes (the ‘between-model’ uncertainty) is indicated by the ‘error bars’. Regions are shown in their approximate north–south and east–west relationship.

Antarctic sector of the Southern Ocean¹³. The p_{CO_2} values south of 50° S showed seasonal variations that require CO_2 uptake in summer and emission in winter. If the seasonality exhibited in this campaign is true for other parts of the Southern Ocean, this would result in a reduction of the Southern Ocean flux uptake in the database, which is currently determined predominantly by summer measurements. This seasonality-driven explanation is also consistent with forthcoming results from the second stage of the TransCom 3 comparison in which we estimate seasonal cycles (K.R.G. *et al.*, manuscript in preparation).

Second, we find carbon uptake over the continents of the Northern Hemisphere to be distributed relatively evenly across North America, Europe and Asia, in contrast to the distribution found in an earlier, widely cited inverse study². We find a temperate North American sink approximately 60% of that found in the earlier study, a small boreal North American source rather than small uptake, and a large sink for Eurasia rather than an approximately neutral flux. Estimated uncertainties are moderate (0.4–0.7 Gt C yr^{-1}), indicating that regional partitioning remains difficult, but the flux differences between the two studies lie at the edge of (or outside) the uncertainty ranges.

Although previous studies have challenged the possibility of a large North American sink^{3–7}, little systematic exploration has been performed as to how such a result was achieved. The differences are not due to the choice of transport model, because the two models used in the earlier study are included here and lie in the middle of our range. Extensive sensitivity tests (see Tables 3 and 4 in Supplementary Information) indicate that the Eurasian flux estimate is

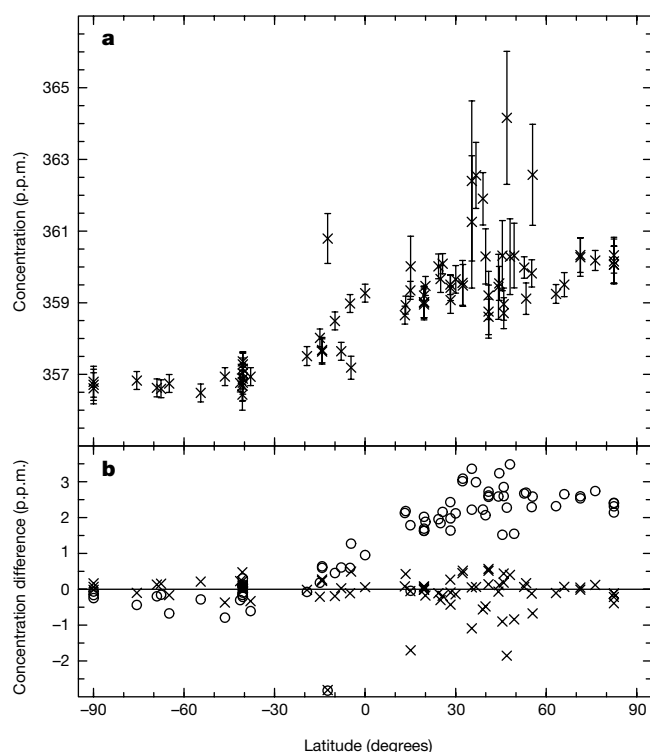


Figure 2 CO_2 concentrations input to, and as fitted by, the inversion. **a**, Meridional gradient of observational CO_2 values and the uncertainty assigned to them in the inversion. **b**, Model mean concentrations for the sum of the three background fluxes minus the observational CO_2 values (circles) and the model mean concentrations after inverting for regional fluxes minus the observational CO_2 values (crosses). Concentrations are 1992 to 1996 means. Note that the adjustment to the background fluxes requires additional uptake in the Northern Hemisphere and lessened uptake in the Southern Hemisphere for all models.

very sensitive to the pattern of background fluxes used in the inversion, especially that representing the seasonal terrestrial biosphere. The difference in North American uptake results from a combination of methodological choices as well as differences in time period and observational stations used.

There are three methodological differences that together appear to be critical. First, recent work¹⁴ suggests that the larger the region size in an inversion, the greater the potential for producing biased flux estimates. Second, the potential bias can be reduced by increasing the data uncertainty for sites in regions with spatially heterogeneous fluxes. The earlier study² inverted for larger regions than used here, and used relatively small (0.6 p.p.m.), spatially invariant uncertainties compared to the generally larger, variable uncertainties used in this study. The third factor is the uncertainty assigned to prior estimates of ocean fluxes, which were zero in the earlier study. Thus the flux adjustment required to match the atmospheric data was applied only to land regions. Together these three factors suggest that the earlier study had greater potential for biased and more sensitive flux estimates than the control results presented here.

Although transport uncertainties do not overwhelm our flux estimates, one factor appears to be responsible for a significant portion of the model spread; the 'rectifier' produced by the covariance between the seasonal biospheric background flux and atmospheric transport¹⁵. The effect of the rectifier can be seen by performing the inversion without the background biospheric fluxes (Fig. 1, right-hand symbols within each box). The between-model uncertainty is reduced for almost all regions, and in some regions there are substantial changes to the estimated fluxes. An increase of 1.1 Gt C yr^{-1} in boreal Asia changes it from a moderate sink to a moderate source, because rectification produces the strongest concentrations downwind of this region in many of the models. Sink strengths increase by 0.35–0.55 Gt C yr^{-1} for temperate North America, temperate Asia and northern Africa, to maintain the required global source. Measurements indicating the strength of the covariance effect in nature are needed to assess this aspect of model transport.

One way to reduce the large uncertainties in our full calculation is to aggregate our regions after performing the inversion. Figure 3 shows the flux estimates for the land and ocean separately in the southern extratropics, tropics, northern extratropics and for the

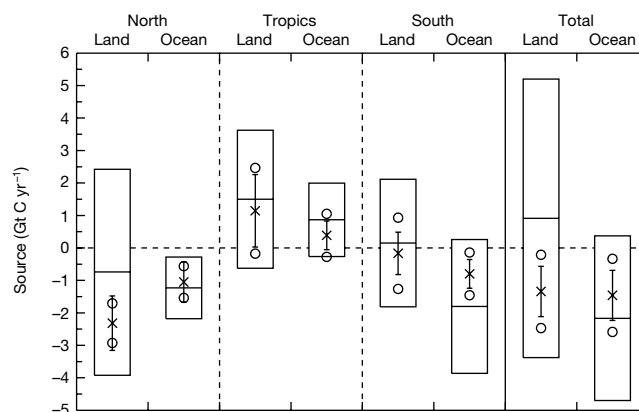


Figure 3 Mean sources and uncertainties for six aggregated regions and global land and ocean. Symbols as in Fig. 1 with the regional aggregation as follows: northern land (boreal North America, temperate North America, Europe, temperate Asia, boreal Asia), tropical land (tropical America, northern Africa, tropical Asia), southern land (South America, southern Africa, Australia), northern ocean (North Pacific, Northern Ocean, North Atlantic), tropical ocean (tropical west Pacific, tropical east Pacific, tropical Atlantic, tropical Indian), southern ocean (Southern Ocean, South Atlantic, South Indian).

globe as a whole. Within-model uncertainties are reduced relative to simply summing from constituent regions, because much of the uncertainty occurs at the scale of the original regions. There is also a reduction in the between-model uncertainty for this aggregation, as some of the model spread involves details of regional transport. At this larger scale, the decrease in the sink in the Southern Ocean and the enhanced sinks over the Northern Hemisphere appear more significant. The uncertainty on the tropical land region is large; lack of atmospheric data in this region means that inversion methods cannot reliably comment on the extent to which sources due to tropical land-use change are balanced by enhanced growth.

This first stage of the TransCom 3 intercomparison has explored many aspects of annual mean inversions more comprehensively than previous work. By incorporating a range of transport models, the fluxes and their uncertainties represent progress towards more robust inverse estimates of regional carbon exchange. Carbon exchange with the ocean is well constrained in this study and, in the case of the Southern Ocean region, is different from fluxes suggested by p_{CO_2} measurements. This result is consistent across the 16 transport models used here, and is insensitive to many aspects of the inversion set-up. Flux estimates in the northern extratropical land regions are reasonably robust as zonal means, but are difficult to distinguish in the longitudinal direction, and can be biased owing to key methodological aspects of the inversion construction. Seasonal exchange with the terrestrial biosphere is responsible for much of the model spread over these regions. Realistic characterization of this aspect of model transport is essential if this uncertainty is to be reduced in the future.

Flux estimates in the tropical land regions remain very uncertain, owing to few CO_2 observations and the limited influence of extratropical observations on the tropical land flux estimates. New observations that can be represented by global-scale transport models are needed in these regions. Future TransCom analysis will focus on the effect of transport model differences on flux estimates at seasonal and interannual timescales. □

Methods

We use a Bayesian synthesis inversion formalism¹⁶ that specifies prior estimates of both the fluxes and their uncertainty, and optimizes with respect to atmospheric observations that are also uncertain. We estimate fluxes for 11 land and 11 ocean regions (see Supplementary Information) as differences from 'background' fluxes that are run separately through each transport model and represent fossil-fuel emissions^{17,18}, seasonally varying air-sea gas exchange¹⁰ and an annually balanced, seasonally varying flux due to terrestrial photosynthesis and respiration¹⁹. The use of seasonally varying background fluxes allows the annual mean inversion to include contributions to annual mean concentrations due to the covariance of atmospheric transport and seasonal fluxes.

We invert 5-year mean measurements for 1992–96 at 76 sites taken from the GLOBALVIEW-2000 data set²⁰ (see Supplementary Information). GLOBALVIEW is a data product that interpolates CO_2 measurements to a common time interval. Gaps in the data are filled by extrapolation from marine boundary layer measurements. We have chosen to use sites where the extrapolated data accounts for less than 30% of the 1992–96 period. The measurements are weighted inversely by the degree to which the predicted concentrations are required by the inverse process to match the observations, which we refer to as 'data uncertainty'. In addition to measurement precision, this uncertainty incorporates the inability of coarse-grid models to adequately represent discrete measurements. The relative uncertainty of one site to another was based on the mean residual standard deviations for 1992–96 from GLOBALVIEW. The absolute magnitudes were chosen to produce a mean square normalized residual out of the inversion of about 1.0, ensuring that the estimated fluxes were optimized to the data only to an appropriate level commensurate with our ability to model them. A minimum uncertainty was also specified. This gave uncertainties ranging from 0.25 p.p.m. for remote, 'clean air' sites to 2.2 p.p.m. for continental, 'noisy' sites (Fig. 2).

The 11 land basis region boundaries were constructed to enclose vegetation of similar seasonal structure and carbon exchange based on vegetation classification²¹. Ocean basis regions were chosen to approximate circulation features such as gyres and upwelling regions. Unit emissions of 1 Gt C yr^{-1} were specified from each region. Subregional-scale variations in emissions rates were prescribed for land regions according to simulated net primary production from the CASA model²¹. This assumes that carbon fluxes follow the distribution of vegetation productivity. Emissions from ocean regions were prescribed as spatially uniform, except that sea-ice was masked out using seasonally varying fractional ice cover distributions²². The inversion requires prior flux and uncertainty estimates. Our choices have been guided by ocean and terrestrial flux models and observations^{10,19}, and are shown in Fig. 1 (also see Table 2 in Supplementary Information). The land region prior

flux estimates incorporate recent inventory estimates^{23–30}. Where more than one estimate for a given region was considered, a mid-point of the estimate spread was used. The prior flux uncertainty was chosen to be large enough to encompass all estimates. Prior flux uncertainties reflect one standard deviation.

The inversion is run separately for 16 transport models or model variants. The models used (and the initials of the modellers) are CSU general circulation model (K.R.G., A.S.D.), Goddard Institute for Space Studies off-line model—UCB (I.Y.F., J.J.), UCI-CTM with GISS-II' fields (M.P., B.C.P., 3 model variants), Japan Meteorological Agency—CDTM (T.M.), MATCH/CCM3 winds (L.B.), MATCH/NCEP winds (Y.H.C.), MATCH/MACCM2 winds (R.M.L.), NIES (S.M.), National Institute for Resources and Environment (S.T.), Recherche en Prevision Numerique (C.W.Y.), SKYHI (S.F.), TM2 (P.B., P.C., P.P.), TM3 (M.H.), GCTM (D.B.). The inversion produces estimated fluxes and their uncertainties for each region individually and for some groups of regions in addition to a background concentration. Here our analysis focuses on fluxes and uncertainties that are averaged across models. We also specify two measures of uncertainty as described in the main text. We show the 'between' model uncertainty as a one standard deviation confidence interval for meaningful comparison with the within model uncertainty. The obvious alternative, showing a full range, would also produce a confidence interval that would widen as more models were included. Inspection of the individual flux estimates showed them to be close to normally distributed about the mean flux for most regions.

Received 14 June; accepted 11 December 2001.

- Enting, I. G., Trudinger, C. M. & Francey, R. J. A synthesis inversion of the concentration and $\delta^{13}\text{C}$ of atmospheric CO_2 . *Tellus B* **47**, 35–52 (1995).
- Fan, S. *et al.* A large terrestrial carbon sink in North America implied by atmospheric and oceanic carbon dioxide data and models. *Science* **282**, 442–446 (1998).
- Kaminski, T., Heimann, M. & Giering, R. A coarse grid three-dimensional global inverse model of the atmospheric transport, 2, inversion of the transport of CO_2 in the 1980s. *J. Geophys. Res.* **104**, 18555–18581 (1999).
- Bousquet, P., Ciais, P., Peylin, P., Ramonet, M. & Monfray, P. Inverse modelling of annual atmospheric CO_2 sources and sinks. Part 1: method and control inversion. *J. Geophys. Res.* **104**, 26161–26193 (1999).
- Baker, D. F. *Sources and Sinks of Atmospheric CO_2 Estimated from Batch Least-Squares Inversions of CO_2 Concentration Measurements*. Thesis, Princeton Univ. (2001).
- Taguchi, S. Synthesis inversion of atmospheric CO_2 using the NIRE chemical transport model. *Geophys. Monogr.* **114**, 239–254 (2000).
- Peylin, P., Baker, D., Sarmiento, J., Ciais, P. & Bousquet, P. Influence of transport uncertainty on annual mean and seasonal inversions of atmospheric CO_2 data. *J. Geophys. Res.* (submitted).
- Rayner, P. J., Enting, I. G., Francey, R. J. & Langenfelds, R. Reconstructing the recent carbon cycle from atmospheric CO_2 , $\delta^{13}\text{C}$ and O_2/N_2 observations. *Tellus B* **51**, 213–232 (1999).
- Bousquet, P. *et al.* Regional changes in carbon dioxide fluxes of land and ocean since 1980. *Science* **290**, 1342–1346 (2000).
- Takahashi, T. *et al.* Global sea-air CO_2 flux based on climatological surface ocean p_{CO_2} , and seasonal biological end temperature effects. *Deep-Sea Res. I.* (submitted).
- Tans, P., Fung, I. & Takahashi, T. Observational constraints on the global atmospheric CO_2 budget. *Science* **247**, 1431–1438 (1990).
- Denning, S. *et al.* Three-dimensional transport and concentration of SF_6 : A model intercomparison study (Transcom 2). *Tellus B* **51**, 266–297 (1999).
- Metz, N., Brunet, C., Jabaud-Jan, A., Poisson, A. & Schauer, B. in *Extended Abstr. 6th Int. Carbon Dioxide Conf.* 685–688 (Organizing Committee of Sixth Carbon Dioxide Conference, Sendai, 2001).
- Kaminski, T., Rayner, P. J., Heimann, M. & Enting, I. G. On aggregation errors in atmospheric transport inversions. *J. Geophys. Res.* **106**, 4703–4715 (2001).
- Denning, A. S., Fung, I. Y. & Randall, D. A. Latitudinal gradient of atmospheric CO_2 due to seasonal exchange with land biota. *Nature* **376**, 240–243 (1995).
- Tarantola, A. *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation* 3rd impr. (Elsevier, Amsterdam, 1998).
- Andres, R. J., Marland, G., Fung, I. & Matthews, E. Distribution of carbon dioxide emissions from fossil fuel consumption and cement manufacture, 1950–1990. *Glob. Biogeochem. Cycles* **10**, 419–429 (1996).
- Brenkert, A. L. Carbon dioxide emission estimates from fossil-fuel burning, hydraulic cement production, and gas flaring for 1995 on a one degree grid cell basis. (<http://cdiac.esd.ornl.gov/ndps/ndp058a.html>) (1998; accessed Oct. 1998).
- Randerson, J. *et al.* The contribution of terrestrial sources and sinks to trends in the seasonal cycle of atmospheric carbon dioxide. *Glob. Biogeochem. Cycles* **11**, 535–560 (1997).
- GLOBALVIEW-CO₂: Cooperative Atmospheric Data Integration Project - Carbon Dioxide CD-ROM (NOAA CMDL, Boulder, Colorado, 2000); also available at (<ftp://ftp.cmdl.noaa.gov/ccg/co2/>) GLOBALVIEW (2000).
- De Fries, R. S. & Townshend, J. R. G. NDVI-derived land cover classifications at a global scale. *Int. J. Remote Sensing* **15**, 3567–3586 (1994).
- Taylor, K. E., Williamson, D. & Zwiers, F. AMIP II sea surface temperature and sea ice concentration boundary conditions. (<http://www-pcmdi.llnl.gov/amip/AMIP2EXPDSN/BCS/amip2bcs.html>) (1997).
- Apps, M. J. & Kurz, W. A. in *Carbon Balance on World's Forested Ecosystems: Towards a Global Assessment* (ed. Kanninen, M.) 14–39 (Publications of the Academy of Finland, Helsinki, 1994).
- Kurz, W. A. & Apps, M. J. A 70 year retrospective analysis of carbon fluxes in the Canadian forest sector. *Ecol. Appl.* **9**, 526–547 (1999).
- Greenhouse Gas Inventory Data from 1990 to 1998 (Secretariat of the United Nations Framework Convention on Climate Change, National Communications from Parties Included in Annex 1 to the Convention, FCCC/SBI/2000/11, Bonn, 2000).
- Pacala, S. *et al.* Convergence of land- and atmosphere-based U.S. carbon sink estimates. *Science* **292**, 2316–2320 (2001).
- Houghton, R. A. The annual net flux of carbon to the atmosphere from changes in land use 1850–1990. *Tellus B* **51**, 298–313 (1999).
- Dixon, R. K. *et al.* Carbon pools and flux of global forest ecosystems. *Science* **263**, 185–190 (1990).

29. Houghton, R. A. & Hackler, J. L. Emissions of carbon from forestry and land-use change in tropical Asia. *Glob. Change Biol.* **5**, 481–492 (1999).
30. Kauppi, P. E., Mielikainen, K. & Kuusela, K. Biomass and carbon budget of European forests, 1971 to 1990. *Science* **256**, 70–74 (1992).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank B. Stephens for comments and suggestions on earlier versions of the manuscript. This work was supported by the NSF, NOAA and the International Geosphere Biosphere Program/Global Analysis, Interpretation, and Modeling Project. S.F. and J.S. were supported by NOAA's Office of Global Programs for the Carbon Modeling Consortium.

Correspondence and requests for materials should be addressed to A.S.D. (e-mail: denning@atmos.colostate.edu).

Unsuspected diversity among marine aerobic anoxygenic phototrophs

Oded Béjà^{*†}, Marcelino T. Suzuki^{*}, John F. Heidelberg[‡], William K. Nelson[‡], Christina M. Preston^{*}, Tohru Hamada^{§†}, Jonathan A. Eisen[‡], Claire M. Fraser[‡] & Edward F. DeLong^{*}

^{*} Monterey Bay Aquarium Research Institute, Moss Landing, California 95039-0628, USA

[‡] The Institute for Genomic Research, Rockville, Maryland 20850, USA

[§] Marine Biotechnology Institute, Kamaishi Laboratories, Kamaishi City, Iwate 026-0001, Japan

Aerobic, anoxygenic, phototrophic bacteria containing bacteriochlorophyll *a* (Bchl*a*) require oxygen for both growth and Bchl*a* synthesis^{1–6}. Recent reports suggest that these bacteria are widely distributed in marine plankton, and that they may account for up to 5% of surface ocean photosynthetic electron transport⁷ and 11% of the total microbial community⁸. Known planktonic anoxygenic phototrophs belong to only a few restricted groups within the Proteobacteria α -subclass. Here we report genomic analyses of the photosynthetic gene content and operon organization in naturally occurring marine bacteria. These photosynthetic gene clusters included some that most closely resembled those of Proteobacteria from the β -subclass, which have never before been observed in marine environments. Furthermore, these photosynthetic genes were broadly distributed in marine plankton, and actively expressed in neritic bacterioplankton assemblages, indicating that the newly identified phototrophs were photosynthetically competent. Our data demonstrate that planktonic bacterial assemblages are not simply composed of one uniform, widespread class of anoxygenic phototrophs, as previously proposed⁸; rather, these assemblages contain multiple, distantly related, photosynthetically active bacterial groups, including some unrelated to known and cultivated types.

Most of the genes required for the formation of bacteriochlorophyll-containing photosystems in aerobic, anoxygenic, phototrophic (AAP) bacteria are clustered in a contiguous, 45-kilobase (kb) chromosomal region (superoperon)⁶. These include *bch* and *crt* genes coding for the enzymes of the bacteriochlorophyll and carotenoid biosynthetic pathways, and the *puf* genes coding for the subunits of the light-harvesting complex (*pufB* and *pufA*) and the reaction centre complex (*pufL* and *pufM*). To better describe the nature and diversity of planktonic, anoxygenic, photosynthetic bacteria, we screened a surface-water bacterial artificial chromo-

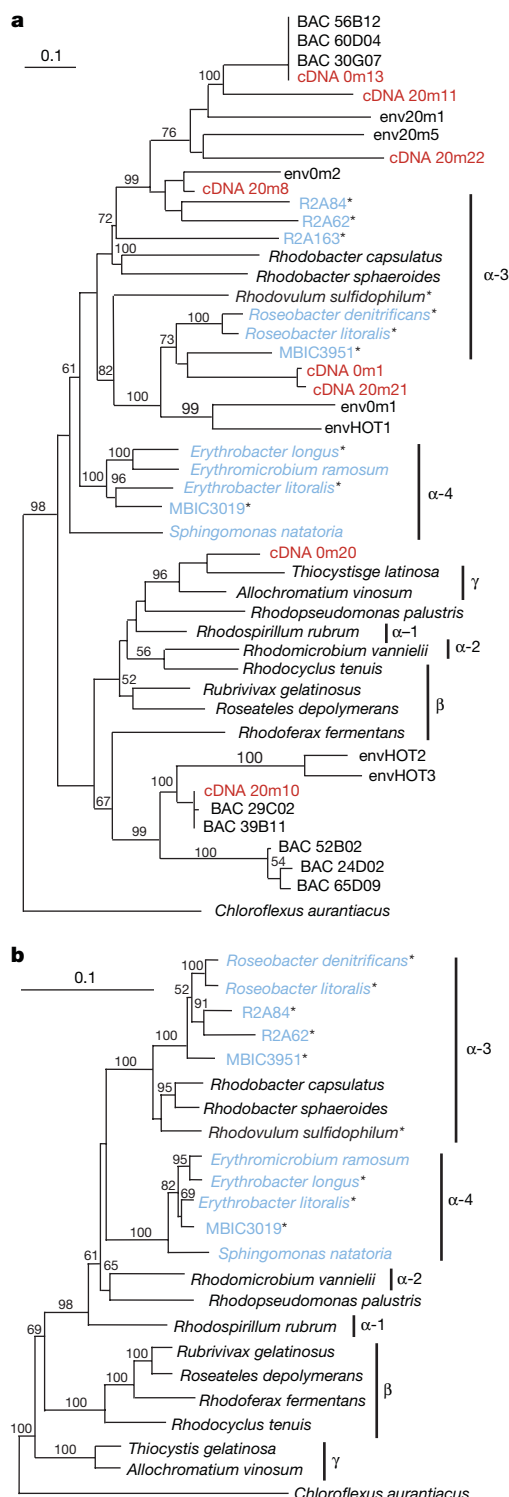


Figure 1 Phylogenetic relationships of *pufM* gene (a) and rRNA (b) sequences of AAP bacteria. **a, b**, Evolutionary distances for the *pufM* genes (a) were determined from an alignment of 600 nucleotide positions, and for rRNA genes (b) from an alignment of 860 nucleotide sequence positions. Evolutionary relationships were determined by neighbour-joining analysis (see Methods). The green non-sulphur bacterium *Chloroflexus aurantiacus* was used as an outgroup. *pufM* genes that were amplified by PCR in this study are indicated by the env prefix, with 'm' indicating Monterey, and HOT indicating Hawaii ocean time series. Cultivated aerobes are marked in light blue, bacteria cultured from sea water are marked with an asterisk, and environmental cDNAs are marked in red. Photosynthetic α -, β - and γ -proteobacterial groups are indicated by the vertical bars to the right of the tree. Bootstrap values greater than 50% are indicated above the branches. The scale bar represents number of substitutions per site.

[†] Present addresses: Department of Biology, Technion-Israel Institute of Technology, Haifa 32000, Israel (O.B.); Nara Institute of Science and Technology, Takayama, Ikoma 630-0101, Japan (T.H.).

some (BAC) library⁹ prepared from surface-water marine bacterioplankton, for clones containing *pufL* and *pufM*^{10,11} genes.

Several *puf*-containing BAC clones were identified in screens using primers¹⁰ designed to amplify almost the entire *pufL* and *pufM* region (1520–1580 nucleotides). In addition, several coastal α -proteobacterial strains previously isolated off the Oregon coast were also screened (R2A strains)¹². The *pufM* phylogenetic tree (Fig. 1a) encompassed two principal clades, one containing α -3 and α -4 proteobacteria and one containing α -1, α -2, β - and γ -proteobacteria representatives, in agreement with an earlier study¹⁰. The BAC clones fell into three groups on the basis of *pufM* gene phylogeny (Fig. 1a). One group (BAC clones 30G07, 56B12 and 60D04) was most closely related to *pufM* from α -proteobacteria isolates R2A62 and R2A84 (ref. 12) (*Roseobacter*-like, Fig. 1b), and was placed in the group containing α -3 Proteobacteria. The other two groups (29C02, 39B11 and 24D02, 52B02, 65D09) branched together, and were most similar to the freshwater β -proteobacterium *Rhodospirillum rubrum* (Fig. 1a, b). *puf* genes were also amplified by polymerase chain reaction (PCR) from DNA extracts of a mixed bacterioplankton assemblage, sampled at the same location as where the BAC library originated: Monterey Bay, California. These Monterey Bay *pufM* sequences clustered with BAC clones 30G07, 56B12 and 60D04, the α -proteobacteria isolates R2A62 and R2A84 (env0m2, env20m1 and env20m5), and the group composed of *Roseobacter* isolates (env0m1). The phylogenetic relationships of *pufM* genes were generally consistent with those determined from ribosomal RNA gene sequences, with the exception of the *pufM* cluster containing the α -1, α -2, β - and γ -proteobacteria (Fig. 1b).

To identify groups actively expressing photosynthetic genes in natural populations, we used PCR with reverse transcription (RT-PCR) to identify photosynthetic-operon messenger RNAs from the same environment. The *pufL/M* primer set failed to amplify any complementary DNA. However, a different primer targeting a

smaller fragment (156 nucleotides) of the *pufM* gene revealed five different groups of *pufM* in a Monterey Bay *pufM* cDNA gene library (cDNAs in Fig. 1). One cDNA group clustered with the aerobic marine phototroph *Roseobacter denitrificans*, another with strain R2A84, and one was most similar to the *pufM* sequence from a γ -proteobacterium, *Thiocystis gelatinosa*. Two *pufM* cDNA groups clustered with the different BAC clones, implying that these BAC inserts originate from bacteria that were actively expressing photosynthetic genes.

To better characterize the bacteria from which these photosynthetic operons originated, the BAC inserts (29C02, 41 kb; 60D04, 103 kb; 65D09, 87 kb) were fully sequenced and the photosynthetic operons compared with those of cultured proteobacteria. The operon organization of BAC clone 65D09 (and 29C02; data not shown), which falls inside the α -1/ α -2/ β / γ *pufM* cluster, is considerably different from the operon organization of any previously reported photosynthetic organism (Fig. 2). Clones 65D09 and 29C02 are similar in their photosynthetic operon organization (data not shown), and when compared with both α - or β -proteobacterial (*Rhodospirillum rubrum*¹³ and *Rubrivivax gelatinosa*¹⁴, respectively) operon organization, more closely resemble the β -proteobacterial operon (Fig. 2). BAC clone 60D04, which is related to α -3 proteobacteria (on the basis of *pufM* sequences), also differs significantly in gene arrangement from any other photosynthetic operon organization reported to date. It most closely resembles the photosynthetic operons from *Rhodospirillum rubrum*⁸ and *R. sphaeroides*^{13,15}, in both organization and gene content (it contains the *bchI*DO and *crtD* genes, which are missing in *R. gelatinosa*¹⁴ and BAC clones 29C02 and 65D09). The superoperon gene arrangement, *crtEF-bchCXYZ-puf* and *bchFNBHLM-lhaA-puhA*^{13–17}, found in *R. sphaeroides*, *R. rubrum* and *R. gelatinosa* are conserved among the naturally occurring planktonic bacterial genomes (Fig. 2), suggesting that the photosynthetic apparatus in the bacteria from which the BACs originated are functional.

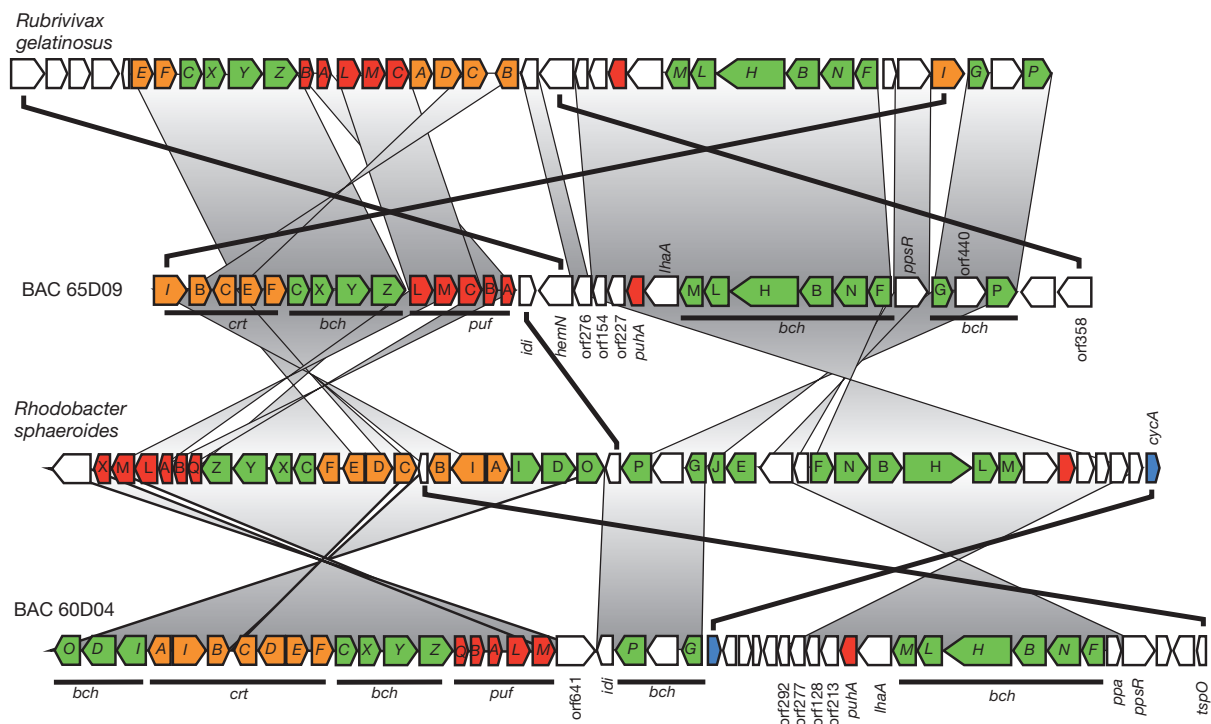


Figure 2 Schematic comparison of photosynthetic operons from *R. gelatinosa* (β -proteobacteria), *R. sphaeroides* (α -proteobacteria) and uncultured environmental BACs. ORF abbreviations use the nomenclature defined in refs 13, 14 and 24. Predicted ORFs are coloured according to biological category: green, bacteriochlorophyll biosynthesis

genes; orange, carotenoid biosynthesis genes; red, light-harvesting and reaction centre genes; and blue, cytochrome *c*₂. White boxes indicate non-photosynthetic and hypothetical proteins with no known function. Homologous regions and genes are connected by shaded vertical areas and lines, respectively.

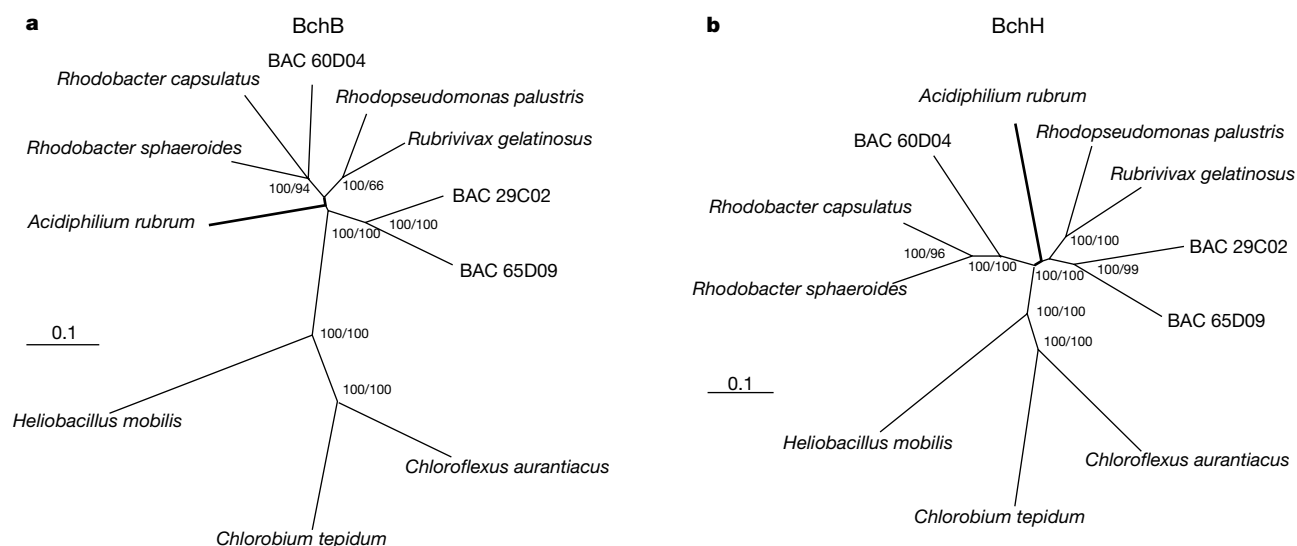


Figure 3 Phylogenetic analyses of BchB and BchH proteins. **a**, Phylogenetic tree for the BchB protein. **b**, Phylogenetic tree for the BchH protein. The BchH sequences from *Chlorobium vibrioforme*²⁵ and BchH2 and BchH3 from *C. tepidum*¹⁸ were omitted from the tree because these genes potentially encode an enzyme for bacteriochlorophyll *c*

We compared several photosynthetic genes found on the oceanic bacteriochlorophyll superoperons with characterized homologues from cultured photosynthetic bacteria. The relationships among Bchl_a biosynthetic proteins BchB (a subunit of light-independent protochlorophyllide reductase) and BchH (magnesium chelatase) were determined¹⁸ (Fig. 3). Similar to *pufM* relationships, BchB and BchH proteins from BAC clone 60D04 were most similar to homologues from *R. capsulatus* and *R. sphaeroides* (Fig. 3). Bchl_a biosynthetic proteins from BAC clones 29C02 and 65D09 were most closely related to those of *R. gelatinosus* (no sequences are yet available from a photosynthetic γ -proteobacterium) and *Rhodospseudomonas palustris*, again in agreement with the phylogenetic relationships of the *pufM* sequences (Fig. 1).

Recent analyses suggest possible horizontal transfer of the photosynthetic gene cluster in purple bacteria¹⁴, and this possibility complicates definitive identification of the organismal origins of these operons based solely on photosynthetic gene analysis. However, gene assignment of open reading frames (ORFs) revealed that more than 75% of ORFs outside the photosynthetic superoperon on BAC 65D09 are most similar to proteins from γ -proteobacteria, whereas ORFs from BAC clone 60D04 are most similar to those of α -proteobacteria. The phylogenetic assignments on the basis of *puf* gene similarities and arrangement are therefore consistent with the chromosomal context external to the respective photosynthetic superoperons in the BAC clones analysed.

In this study, no sequences recovered in Monterey Bay waters were similar to those of *Erythrobacter* species—one of the more commonly cultured Bchl_a-containing bacteria recovered from the open ocean⁸. We therefore used the same *pufM* primers on bacterioplankton DNA extracts from waters of the central North Pacific Ocean (Hawaii ocean time series station¹⁹; envHOT clones in Fig. 1). One HOT environmental *pufM* group (represented by envHOT1 clone) clustered with sequences from α -3 proteobacteria isolates, and another group (envHOT2 and envHOT3) clustered with the BAC sequences related to the freshwater β -proteobacterium, *R. fermentans*. The results suggest that similar groups involved in oceanic aerobic anoxygenic photosynthesis are found in surface waters from both neritic and oceanic systems. These groups do not appear to be similar (at least with respect to their photosynthetic

biosynthesis and are probably of distinct origin (J. Xiong, personal communication). Bootstrap values (neighbour-joining/parsimony method) greater than 50% are indicated next to the branches. The scale bar represents number of substitutions per site. The position of *Acidiphilium rubrum* (bold branch) was not well resolved by both methods.

operon) to cultivated *Erythrobacter* species.

Recently, it has been suggested that cultivated *Erythrobacter* species (α -4 subclass of proteobacteria) may represent the predominant AAPs in the upper ocean^{8,20}. Surprisingly, we were not able to retrieve photosynthetic operon genes belonging to this group in any of the samples analysed. Furthermore, very few sequences related to *Erythrobacter* species have been reported in 16S rDNA clone libraries constructed from marine plankton DNA, also suggesting that this group may not represent the predominant AAPs in the upper ocean. Some of the representative AAPs we found were related to cultured marine bacteria (*Roseobacter* and *Roseobacter*-like bacteria within the α -3 subclass of proteobacteria). Members of this group have also been frequently retrieved in 16S rRNA clone libraries and represent a large proportion of bacterioplankton rDNAs in coastal waters²¹. However, other groups (found in both BAC and cDNA libraries) were only distantly related to known anoxygenic phototrophs, and have never before been observed in marine plankton. This implies that in addition to *Erythrobacter* and *Roseobacter* species, other yet-to-be-cultivated bacteria (most likely related to β - or γ -proteobacteria) actively participate in oceanic aerobic, bacteriochlorophyll-based photosynthesis. The discovery of new marine AAP bacteria through culture-independent genomic analyses emphasizes the complementary nature of culture-based and cultivation-independent approaches, which taken together provide a much more comprehensive perspective than either does alone. Our new observations should help direct current efforts aimed at characterizing those microbes responsible for oceanic, bacteriochlorophyll-mediated photosynthesis, a newly recognized^{7,8} but poorly understood process in marine plankton. □

Methods

BAC library and environmental cDNA preparation

The surface-water BAC library construction has been described previously⁹ and was prepared from sea water from station M2 (located approximately 45 km offshore of Moss Landing, California) pre-filtered through a GF/A glass fibre filter (approximate particle size less than 1.6 μ m) to remove cell aggregates and larger eukaryotic phytoplankton cells. Sea water from the HOT station (22.4° N, 158.0° W) was collected on 17 March 1998. For preparation of environmental cDNA, sea water was collected from 0 m and 20 m on 26 April 2000 at station 4B off Moss Landing, California (36.77° N, 122.02° W) aboard the RV *Western Flyer*, and picoplankton were collected onto Sterivex Cartridges (Millipore) and

stored as described previously²¹. After the initial treatment with proteinase K and SDS described by ref. 22, total RNA from 100 µl of the lysate from the Sterivex cartridges were further purified using the RNeasy tissue kit (Qiagen) and the protocol for lysed cells. To further remove DNA from the RNA preparations, the samples were treated by the DNasefree kit (Ambion) following the manufacturer's protocol. cDNAs were synthesized by reverse transcription using 2 µl of the RNA extracts using random hexamers as primers, using the TaqMan RT kit (Applied Biosystems) according to the manufacturer's protocol. Genomic DNA contamination of cDNA preparations was examined by 5'-nuclease assays comparing gene copies in cDNA preparations and in controls with no reverse transcriptase added. These control assays tested for the presence of both rRNA²¹ and protein-encoding genes (M. T. Suzuki *et al.*, unpublished results). In both assays we observed no signal in controls without reverse transcriptase.

The annotated BAC insert sequences are available at the Monterey Bay Coastal Ocean Microbial Observatory site at <http://www.tigr.org/tdb/MBMO/>.

pufl and pufM amplification

Primers used in this study were pufl, forward (5'-CTKTCGACTTCTGGTSGG-3')¹⁰; pufM, reverse (5'-CCATSGTCAGCGCCAGAA-3') (a modification of the primer reported by ref. 10); and pufM, forward (5'-TACGGSAACTGTWCTAC-3').

Phylogenetic analysis

pufM sequences from the current study combined with sequences from public databases were translated and the protein sequences aligned using the pileup program of the Wisconsin package (GCG). DNA sequences and protein alignments were imported into a database using ARB software (<http://www.arb-home.de>). We aligned DNA sequences based on the protein alignment. Evolutionary distances were calculated with the dnadist program of the PHYLIP package²³, using the Kimura 2-parameter model. Phylogenetic trees were inferred using the neighbour program of the PHYLIP package. To evaluate the reliability of the branching patterns, 100 random bootstrap re-samplings were performed using the program seqboot, with subsequent phylogenetic analyses performed as above. Ribosomal RNA phylogenetic analyses were performed as described above, on alignments encompassing 860 nucleotide sequence positions. For analysis of the shorter pufM cDNA sequences obtained through RT-PCR, a neighbour-joining tree was imported into the ARB database, and the cDNAs were added to tree using the ARB_PARSIMONY program, without local optimization and using a mask that included only those positions encompassing the cDNA sequences.

Received 23 August; accepted 4 December 2001.

- Harashima, K., Shiba, T. & Murata, N. *Aerobic Photosynthetic Bacteria* (Springer, Berlin, 1989).
- Shiba, T., Simidu, U. & Taga, N. Distribution of aerobic bacteria which contain bacteriochlorophyll *a*. *Appl. Environ. Microbiol.* **38**, 43–48 (1979).
- Yurkov, V. V. & Beatty, J. T. Aerobic anoxygenic phototrophic bacteria. *Microbiol. Mol. Biol. Rev.* **62**, 695–724 (1998).
- Shiba, T., Shioi, Y., Takamiya, K., Sutton, D. C. & Wilkinson, C. R. Distribution and physiology of aerobic bacteria containing bacteriochlorophyll *a* on the east and west coast of Australia. *Appl. Environ. Microbiol.* **57**, 295–300 (1991).
- Yurkov, V. V. *et al.* Phylogenetic positions of novel aerobic, bacteriochlorophyll *a*-containing bacteria and description of *Roseococcus thiosulfatophilus* gen. nov., sp. nov., *Erythromicrobium ramosum* gen. nov., sp. nov., and *Erythrobacter litoralis* sp. nov. *Int. J. Syst. Bacteriol.* **44**, 427–434 (1994).
- Alberti, M., Burke, D. H. & Hearst, J. E. In *Anoxygenic Photosynthetic Bacteria* (eds Blankenship, R. E., Madigan, M. & Bauer, C. E.) 1083–1106 (Kluwer, Dordrecht, The Netherlands, 1995).
- Kolber, Z. S., Van Dover, C. L., Niderman, R. A. & Falkowski, P. G. Bacterial photosynthesis in surface waters of the open ocean. *Nature* **407**, 177–179 (2000).
- Kolber, Z. S. *et al.* Contribution of aerobic phototrophic bacteria to the carbon cycle in the ocean. *Science* **292**, 2492–2495 (2001).
- Béja, O. *et al.* Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ. Microbiol.* **2**, 516–529 (2000).
- Nagashima, K. V. P., Hiraishi, A., Shimada, K. & Matsuura, K. Horizontal transfer of genes coding for the photosynthetic reaction centers of purple bacteria. *J. Mol. Evol.* **45**, 131–136 (1997).
- Achenbach, L. A., Carey, J. & Madigan, M. T. Photosynthetic and phylogenetic primers for detection of anoxygenic phototrophs in natural environments. *Appl. Environ. Microbiol.* **67**, 2922–2926 (2001).
- Suzuki, M. T. *et al.* Bacterial diversity among small-subunit rRNA gene clones and cellular isolates from the same seawater sample. *Appl. Environ. Microbiol.* **63**, 983–989 (1997).
- Choudhary, M. & Kaplan, S. DNA sequence analysis of the photosynthesis region of *Rhodospirillum rubrum* 2.4.1^T. *Nucleic Acids Res.* **28**, 862–867 (2000).
- Igarashi, N. *et al.* Horizontal transfer of the photosynthetic gene cluster and operon re-arrangement in purple bacteria. *J. Mol. Evol.* **52**, 333–341 (2001).
- Naylor, G. W., Adlesee, H. A., Gibson, L. C. D. & Hunter, C. N. H. The photosynthesis gene cluster of *Rhodospirillum rubrum*. *Photosyn. Res.* **62**, 121–139 (1999).
- Burke, D. H., Alberti, M. & Hearst, J. E. *bchN* bacteriochlorophyll synthesis genes of *Rhodospirillum rubrum* and identification of the third subunit of light-independent protochlorophyllide reductase in bacteria and plants. *J. Bacteriol.* **175**, 2414–2422 (1993).
- Bauer, C. E., Buggy, J. J., Yang, Z. M. & Marrs, B. L. The superoperon organization of genes for pigment biosynthesis and reaction center proteins is a conserved feature in *Rhodospirillum rubrum*: analysis of overlapping *bchB* and *pufA* transcripts. *Mol. Gen. Genet.* **228**, 433–444 (1991).
- Xiong, J., Fischer, W. M., Inoue, K., Nakahara, M. & Bauer, C. E. Molecular evidence for the early evolution of photosynthesis. *Science* **289**, 1724–1730 (2000).
- Karl, D. M. & Lukas, R. The Hawaii ocean time-series (HOT) program: background, rationale and field implementation. *Deep-Sea Res.* **43**, 129–156 (1996).
- Fenchel, T. Marine bugs and carbon flow. *Science* **292**, 2444–2445 (2001).
- Suzuki, M. T., Preston, C. M., Chavez, F. P. & DeLong, E. F. Quantitative mapping of bacterioplankton populations in seawater: field tests across an upwelling plume in Monterey Bay. *Aquat. Microbiol. Ecol.* **24**, 117–127 (2001).

- Massana, R., Murray, A. E., Preston, C. M. & DeLong, E. D. Vertical distribution and phylogenetic characterization of marine planktonic *Archaea* in the Santa Barbara channel. *Appl. Environ. Microbiol.* **63**, 50–56 (1997).
- Felsenstein, J. PHYLIP-phylogeny interference package (version 3.5). *Cladistics* **5**, 164–166 (1989).
- Hahn, F. M., Baker, J. A. & Poulter, C. D. Open reading frame 176 in the photosynthesis gene cluster of *Rhodospirillum rubrum* encodes *idi*, a gene for isopentenyl diphosphate isomerase. *J. Bacteriol.* **178**, 619–624 (1996).
- Petersen, B. L., Moller, M. G., Stumm, B. M. & Henningsen, K. W. Structure and organization of a 25 kbp region of the genome of the photosynthetic green sulfur bacterium *Chlorobium vibrioforme* containing Mg-chelatase encoding genes. *Hereditas* **129**, 131–142 (1998).

Acknowledgements

We thank S. Haddock for computational advice; J. Zehr and D. Karl for the HOT samples; the captains and crew of the RV *Point Lobos* and *Western Flyer* for assistance at sea; and M. Heaney, V. Sapiro, S. Lo, and M. Holmes for database and software support. Preliminary sequence data for *R. palustris* was obtained from the DOE Joint Genome Institute (JGI) at http://www.jgi.doe.gov/JGI_microbial/html/index.html. This research was supported by the National Science Foundation (J.E.H., J.A.E. and E.F.D.) and the David and Lucile Packard Foundation (E.F.D.).

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.F.D. (e-mail: delong@mbari.org). The sequences reported in this study are deposited with GenBank under accession numbers AB018690, AB027515, AF393983–AF393991, AF393993–AF394002, AY044244–AY044250 and AE008919–AE008921.

Sex differences in emigration and mortality affect optimal management of deer populations

T. H. Clutton-Brock*, T. N. Coulson*, E. J. Milner-Gulland†, D. Thomson* & H. M. Armstrong‡

* Large Animal Research Group, Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

† TH Huxley, School of Environment, Earth Science, and Engineering Royal School of Mines, Imperial College, London SW7 2BP, UK

‡ Scottish Natural Heritage, 12 Hope Terrace, Edinburgh EH6 5NP, UK

Populations of red deer that are limited by food, like those of many other ungulates^{1–3}, commonly include more females than males^{4–7}. We assessed the contribution of variation in sex- and age-specific rates of mortality and emigration to density-dependent changes in the adult sex ratio, using long-term observations and demographic experiments involving the red deer population on Rum, Scotland^{4,5}. We incorporated these effects in a stochastic model of local populations under different management regimes to show here that, when female numbers are allowed to increase to more than 60% of the ecological carrying capacity, the sustainable annual harvest of males from local deer populations will fall. Because males are typically culled by fee-paying hunters and generate more income than females^{5,8,9}, income will decrease as the male harvest falls. Because numbers of female deer throughout much of the Highlands probably exceed the threshold at which male density starts to be affected⁵, many managers might be able to raise income from local deer populations by reducing female numbers, with potential benefits to the vegetation of Scottish Highland environments¹⁰.

Populations of red deer (*Cervus elaphus* L.) throughout the Scottish Highlands include roughly twice as many females aged ≥ 1 yr as males (Fig. 1a). To investigate the consequences of increasing population density for males and females, in 1972 we released the red deer population of the 12-km² north block of Rum

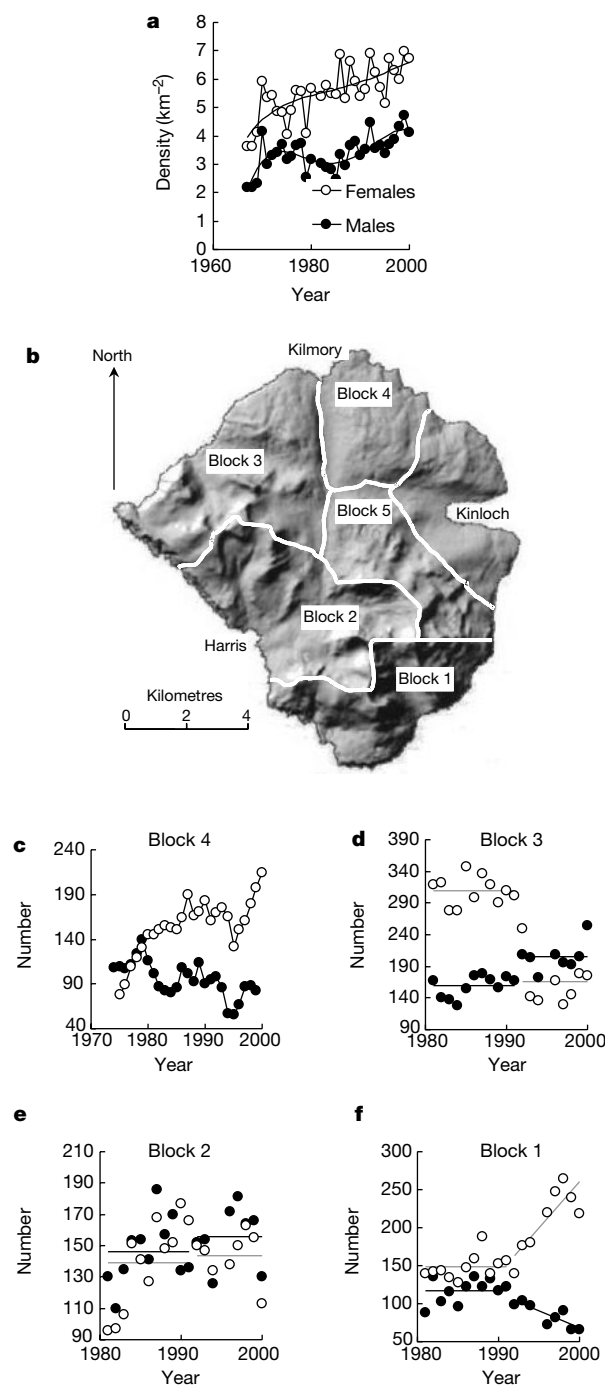


Figure 1 Changes in the adult sex ratio. **a**, Average density of male and females per km² aged ≥ 1 yr across the Scottish Highlands. **b**, Rum showing the outlines of the five deer blocks. **c**, Numbers of identifiable deer aged ≥ 1 yr using the 12-km² north block of Rum (block 4) after the termination of the cull in 1972 (open circles, females; filled circles, males). **d**, Numbers of male and female deer aged ≥ 1 yr during spring counts of the west block of Rum (block 3), before and after high culling rates began to reduce female numbers in 1991–92. Between 1992 and 2000, male numbers were higher than between 1981 and 1991 ($\chi^2 = 21.1$, d.f. = 1, $P < 0.001$). **e**, Numbers of male and female deer using the southwest block of Rum (block 2), where a 14% cull of both sexes was maintained between 1980 and 2000. No significant differences between 1981 and 1992 were present either for males ($\chi^2 = 1.2$, d.f. = 1, $P > 0.05$) or females ($\chi^2 = 1.3$, d.f. = 1, $P > 0.05$). **f**, Numbers of male and female deer using the southeast block (block 1) of Rum before and after high culling rates began to reduce male numbers between 1991 and 1995. After 1991, female numbers showed a progressive increase and were significantly higher between 1992 and 2000 than between 1981 and 1991 ($\chi^2 = 22.5$, d.f. = 1, $P < 0.001$). Methods were as in **d**. Few deer were resident in block 5, which comprised higher ground.

(block 4, Fig. 1b) from the annual 14% cull that had been used to control deer density since 1958^{5,11,12} (Methods). Subsequently, female numbers in block 4 increased by 100%, whereas male numbers gradually decreased (ref. 6; and Fig. 1c).

Five different density-dependent processes contributed to changes in the adult sex ratio in block 4. As the total number of deer rose, the proportion of males born decreased (ref. 13; and Fig. 2a). Neonatal mortality did not change with density (Fig. 2b), but the mortality of males in their first 2 yr increased relative to females of the same age (ref. 14; and Fig. 2c). Mortality of adults (5–10 yr) increased in both sexes, but was higher in males than in females (Fig. 2d). The proportion of males emigrating to other subpopulations (usually aged 2–4 yr; Methods) rose (Fig. 2e) and the number of males immigrating into the population decreased (Fig. 2f), whereas emigration and immigration by females was rare irrespective of density¹⁵.

By calculating the effects on the adult sex ratio when each of these changes was omitted in turn, we identified changes in male emigration and male survival to be the principal causes of the relative increase in female numbers with rising density (Table 1). The antler size of mature males also showed a progressive decrease after 1972 (refs 4, 5). Increasing emigration probably contributed to this, because better grown males were more likely to emigrate as adolescents than smaller individuals (Table 2), and well-grown adolescents carry relatively large antlers as adults^{4,5,15}.

If uncontrolled increases in female deer numbers generate biases in the adult sex ratio, then reductions in female density should lead to increases in male numbers. To determine whether this is the case, we selected the west block of Rum (block 3) (which had previously contained more females than males; see Fig. 1d) and increased the annual cull of females from 14% to ~60% in 1991 and 1992, until female numbers were reduced by 50% (see Fig. 1d). Roughly the same number of adult males (20) continued to be culled each year from the block, while after 1992 the annual 14% cull of females was re-imposed. After the reduction in female density, the number of resident males in block 3 increased by ~25% and is still increasing (Fig. 1d). Over the same period, no significant increase in male numbers occurred in either of the two neighbouring blocks (2 and 4), where management regimes remained unchanged throughout this period (Fig. 1e, c).

To determine whether reductions in male density had similar effects on female numbers, we reduced male numbers by 50% in the southeast block of Rum (block 1, Fig. 1f), but maintained the same percentage cull on females as before 1991. Low male densities immediately after the cull attracted young males from neighbouring areas and male numbers proved difficult to reduce, despite an annual average cull of 68% of the spring population of males each year between 1991 and 1999. By 1995, however, male numbers in block 1 had decreased by 50% (Fig. 1f) and female numbers in the

Table 1 Effects of demographic changes on adult sex ratios

	Adult sex ratio (% males)	Absolute difference
Observed sex ratio in 1973	55.7	
Observed sex ratio in 1982	34.7	
Predicted sex ratios in 1982 assuming		
no density dependence in % males born	36.2	1.5
no differential effect of density on calf winter survival	38.4	3.7
no differential effect of density on yearling survival	38.1	3.4
no differential effect of density on adult survival	43.2	8.5
no differential effect of density on emigration	44.1	9.4
no differential effect of density on immigration	35.9	1.2

The table shows the contributions of six demographic processes to changes in the adult sex ratio in the north block of Rum between 1973 and 1982. The lower part of column one shows predicted adult sex ratios in 1982, calculated assuming that each demographic change had not occurred. Column two shows the absolute difference between the observed sex ratio in 1982 and the predicted value and provides an index of the relative importance of changes in each demographic variable.

block showed a consistent increase, rising by over 50% by 2000. Because rates of female immigration are usually low (Fig. 1e, f), this change was probably a consequence of increased local recruitment resulting from lower population density. The similar responses of male and female numbers to reduced density of the opposite sex indicate that, despite differences in habitat use between the sexes^{16–18}, similar resources probably limit male and female numbers.

Differences in density-dependent responses between male and female deer will have local economic consequences. As in many other parts of Europe, most Scottish red deer populations are managed by local landowners or their representatives, and culling regimes commonly differ between neighbouring land units^{5,19}. In many parts of the Highlands, female culling rates have been lower than the annual rate of recruitment for several decades⁵ though they have recently increased²⁰. As a result, female numbers in many areas are probably above the threshold at which they depress male density, contributing (with other factors, including higher culling rates imposed on males^{5,21}) to the pervasive female bias in the adult sex ratio⁵ (see Fig. 1a).

Income derived from culling males is usually higher than from culling females⁹. Precise figures are unavailable, but a conservative estimate would be a net gain of £200 per adult male culled versus a small net loss per female culled. Female-biased adult sex ratios are consequently likely to be associated with a reduction in potential income relative to deer density. To investigate the implications of density-dependent changes in male and female numbers, we constructed a stochastic model of local red deer populations on Rum under different culling regimes, incorporating the density-dependent changes in sex-specific mortality and emigration shown in Fig. 2 (Methods). Our model investigates the effects of varying culling rates in two neighbouring populations, assuming that no males are culled before they are 5 yr old and that female numbers rise to ecological carrying capacity (k) if no females are culled. In our model, k corresponds to a predicted density of roughly 20 females per km² (Fig. 3a), which is similar to observed densities in the north block of Rum; however, the absolute value of k does not have a qualitative effect on our conclusion (Methods).

In our model, female density decreases as the percentage of females culled increases and is higher when neighbouring female populations are not culled (and female emigration rates are consequently low) than when 10% of females in neighbouring areas are culled each year (Fig. 3a). The density of mature males, as well as the potential annual off-take, is higher when few males are culled each year than when a high proportion of males are removed annually (Fig. 3b). For all male culling rates, the highest densities and potential off-takes of males are achieved in populations where

female culling rates are relatively high and female density is relatively low (thus minimizing male emigration and mortality). For any particular culling rate of females, male densities are higher where females in neighbouring areas are not culled or are lightly culled (leading to relatively high female densities and low rates of male immigration) than in areas where neighbouring female populations are subject to higher culling rates and female density is lower (Fig. 3b, c).

Male harvests, in our model, are maximized at relatively low female densities. If neighbouring female populations are not culled, then maximum male densities will be achieved if female populations are reduced to ~50% of the ecological carrying capacity; by contrast, if 10% of females are culled in neighbouring areas (and male emigration rates are higher), then maximum male densities will be achieved if female density is maintained at ~60% of the ecological carrying capacity. These results remain essentially unchanged if male density is replaced by estimated income from management, assuming a return of £200 per male culled and a zero return per female (Fig. 3d, e). Changes in relative income from males and females (ranging from an income of £50 for individuals of each sex to a gain of £250 per male and a loss of £50 per female) have little effect on the optimal density of females.

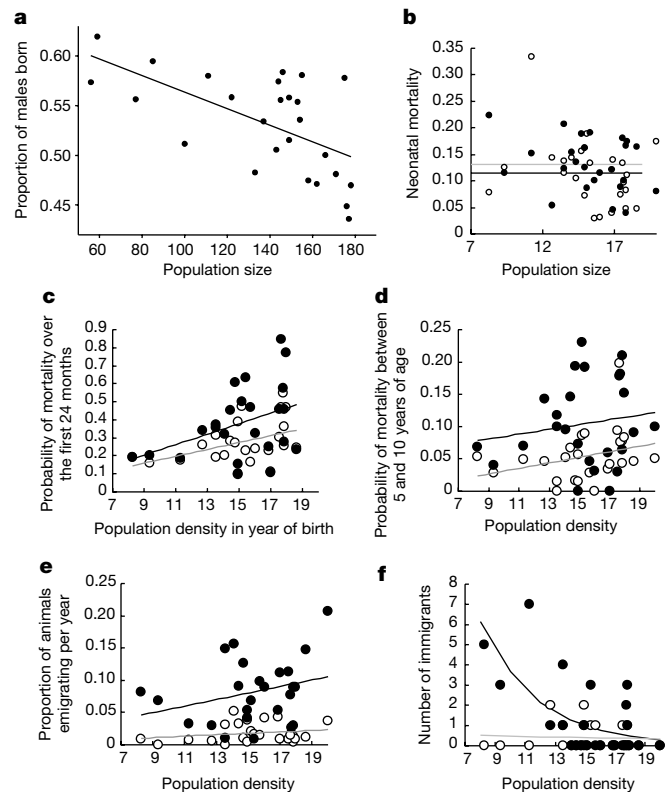


Figure 2 Demographic parameters contributing to the change in adult sex ratio in the deer population of the north block after 1972 (see Fig. 1c). Open circles, females; filled circles, males. **a**, Percentage of calves born each year that were male plotted on population size (males and females), after effects of winter weather were controlled for (from ref. 13). **b**, Probability of mortality of male and female calves in the first three months of life. Mortality over this period was not density-dependent in either sex^{4,5}. **c**, Probability of mortality of males and females in the first 24 months of life (males, $\chi^2 = 3.46$, d.f. = 1, $P < 0.1$; females, $\chi^2 = 7.14$, d.f. = 1, $P < 0.05$)¹⁴. **d**, Probability of mortality of males and females among 5–10-year-olds (males, $\chi^2 = 0.12$, d.f. = 1, $P > 0.05$; females, $\chi^2 = 4.48$, d.f. = 1, $P < 0.05$). **e**, Emigration rates (proportion of residents permanently leaving per year¹⁵) of males and females (males, $\chi^2 = 4.6$, d.f. = 1, $P < 0.05$; females, $\chi^2 = 0.8$, d.f. = 1, $P > 0.05$). **f**, Immigration of males and females (number of permanent immigrants per year¹⁵) (males, $\chi^2 = 11.8$, d.f. = 1, $P < 0.01$; females, $\chi^2 = 0.12$, d.f. = 1, $P > 0.05$).

Table 2 Factors affecting male emigration rates

Term	Estimate	χ^2	P
Fitted alone			
Spike length	0.1227	5.7	< 0.05
Birth weight	0.3007	10.5	< 0.001
Population density	0.01227	6.4	< 0.05
NAO	0.2069	18.2	< 0.001
Full model			
Spike length	0.0979	3.1	< 0.1
Birth weight	0.2799	8.2	< 0.01
Population density	0.0139	7.5	< 0.01
NAO	0.2142	18.5	< 0.001
Regression (d.f.)	4	38.8	
Error (d.f.)	331	421.8	

The table shows the estimated regression values of different variables on the probability that males will disperse from the north block of Rum, derived from a logistic regression model. Measures of population density and the NAO were for the year when the animal was a yearling. Yearlings with longer spikes were more likely to emigrate than those with shorter spikes; males with higher birth weights were more likely to emigrate than males that were lighter at birth; and dispersal rates increased with population density as well as in wet stormy winters when values of the North Atlantic Oscillation (NAO) were high. Adult males with longer spikes as yearlings were heavier and had larger antlers as adults^{4,5}. Data are for 1974–99.

Relationships between male and female density may be affected by a number of environmental variables not included in our model, including the availability of shelter^{22,23}, competition for resources with other herbivores^{24,25}, the size of the management units⁵ and the relative rates of culling imposed on the two sexes. In addition, in

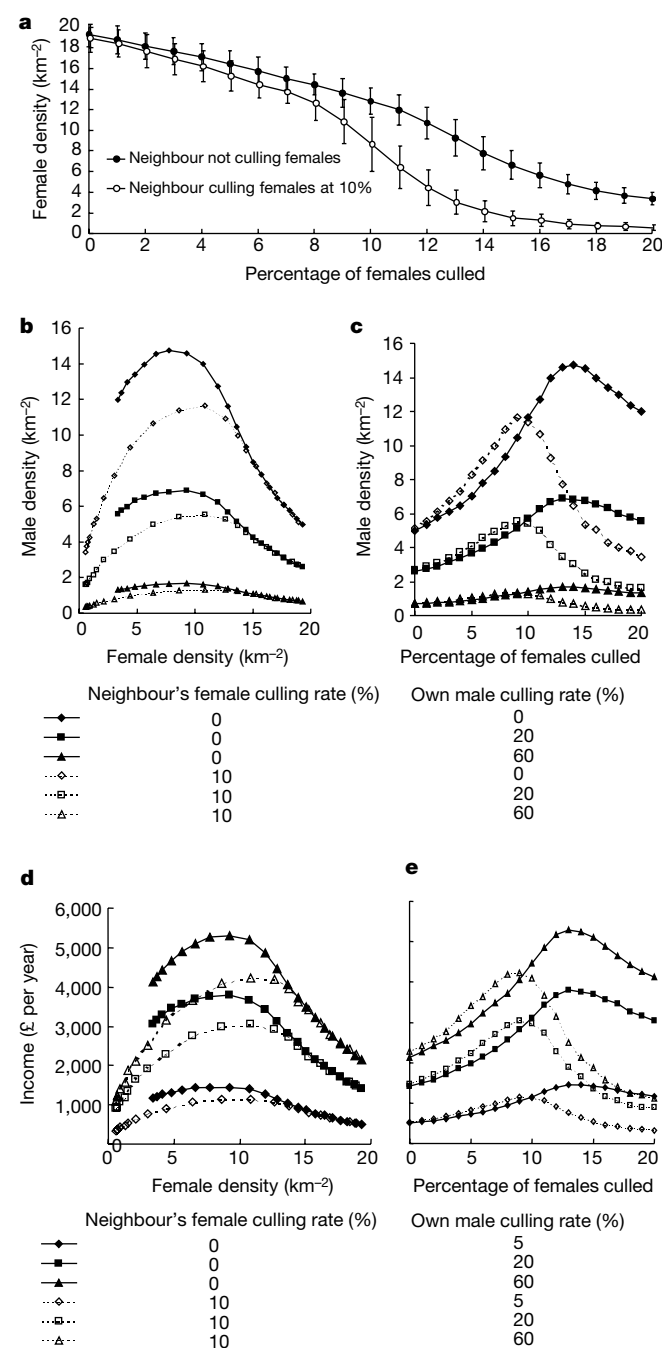


Figure 3 Results of a stochastic model of two contiguous deer populations subject to different culling regimes, incorporating density-dependent changes in age- and sex-specific rates of fecundity mortality, immigration and emigration, derived from the north block of Rum (see Fig. 2). **a**, Estimated equilibrium density of adult females subjected to different levels of annual cull. **b**, Variation in the density of mature males (≥ 5 yr) at different levels of female density, and different annual culling rates of males (varying from 0 to 60%) where females on neighbouring ground are culled, at 0% per year or 10% per year. **c**, Variation in the density of mature males at different levels of female culling rate (0–20% per year). **d**, Estimated total revenue from culling, with returns set at £200 per male culled and zero pounds per female culled, at different female densities. **e**, Estimated total revenue from culling (as above) at different culling rates for females.

some parts of Scotland, mature males leave their usual ranges at the start of the rut and travel to other land units where they may be culled before they return^{5,26}. If high female densities attract temporary 'rut' immigrants that are likely to be culled, such 'honeypot' effects could reduce the benefits of lowering female density. These effects are unlikely to be important where land units are sufficiently large that temporary 'rut' immigrants constitute a small proportion of males culled or where most males do not travel far to rut¹⁵. However, where temporary immigrants represent a substantial proportion of males culled, high female densities could be necessary to maintain the harvest of temporary immigrants, although reductions in female density might be expected to raise the number of resident males and to reduce the number of immigrants¹⁵.

If, as we suggest, female deer numbers in many parts of the Scottish Highlands are substantially above 50% of ecological carrying capacity⁵, by reducing female numbers, many Scottish deer managers may be able not only to reduce female numbers with little risk of affecting the size of male harvests²⁷ but also to increase both their annual off-take of mature males and their net income from deer management. Where managers fail to maintain female numbers below ecological carrying capacity, neighbouring estates that cull females at higher rates and attract adolescent males from adjoining areas, are likely to be the principal beneficiaries (see Fig. 3). Because browsing by red deer is implicated in the loss of heather cover and the prevention of tree regeneration^{10,22,28}, reductions in female numbers are also likely to have benefits for the environment that may, eventually, have further benefits for deer populations. Similar relationships between male and female density probably occur in other sexually dimorphic ungulates where increasing density affects rates of emigration and mortality to a greater extent in males than in females^{29–31}.

Methods

Sex ratios throughout the Highlands

We based estimates of average densities of males and females throughout the Highlands on counts by the Deer Commission (Scotland) at irregular intervals, controlling for the identity of the areas counted⁵.

Demographic consequences of increasing density

Until 1972, red deer populations in all five blocks of Rum had been maintained at a density of about 14 deer per km^2 by an annual cull of about 14% of the spring population^{5,11,12}, but in 1972 the annual cull⁴ in block 4 was terminated and numbers were allowed to increase. Records of changes in deer density in the north block after 1972 were based on five censuses of the area each month between January and April, in which the identity and location of each deer seen was recorded^{4,5}. Individuals were regarded as resident in the block if they were seen in at least 10% of censuses¹⁵. Emigrants were individuals previously resident in the block that left the population, whereas immigrants were those that joined the population after being resident in other parts of the island. Emigration and immigration was commonest in males aged between 2 and 5 yr.

To estimate the contributions of different demographic variables to density-dependent changes in the adult sex ratio, we removed the effects of density on each variable in turn while keeping all other vital rates as observed (Table 1), with the population structure in 1973 providing the initial starting values. Values in Table 1 that are closest to the observed sex ratio in 1982 (34.7% male) represent sex-specific density responses that have little effect on the adult sex ratio. Data used in this analysis were restricted to the period 1973–82 because by 1982 the adult sex ratio had stabilized. Estimates of the effects of birth weight, spike length and climate on emigration (Table 2) were derived from a generalized linear model, using data for 1974–99.

Demographic experiments

To investigate the effects of female density on male numbers and of male density on female numbers, we manipulated the relative densities of females in block 3 and of males in block 1 by altering the annual culling regime. Between 1991 and 1992, female numbers in block 3 were reduced by 50%, and between 1991 and 1995 male numbers in block 1 were reduced by a similar proportion. Annual culls of the sex whose density was not altered (males in block 3 and females in block 1) were maintained during the period of the experiment at the same absolute level as over the past decade. Numbers of deer in each block were counted in May each year by an experienced counting team from the Deer Commission (Scotland) using established techniques^{5,20}.

Population dynamics and density

To assess the likely effects of rising female numbers on the potential harvest of males, we

constructed a stochastic population model, parameterized using the density-dependent relationships shown in Fig. 2. Our model simultaneously considered male and female densities and potential harvests in two neighbouring populations that were small enough for differences in the culling regimes imposed on each population to produce local differences in sex-specific emigration and immigration, as well as in mortality. We assumed that both populations were biologically identical and were subject to correlated environmental variability: relaxation of these assumptions was unlikely to affect the direction of predicted trends. Vital rates used were initially calculated deterministically as $R_j = 1/[1 + \exp(-a + bD)]$, where D was the density of mature females (age classes 3 and above), j was the age class and a and b were constants. Sex ratio at birth was calculated as $m = 64.3 - 0.748D$, where m is the percentage of males. Vital rates were subsequently altered to include variability and correlations as follows: for emigration rates, sex ratio at birth and female mortality, $p_i = R_j + z\sigma_j$; for fecundity and male mortality, $\rho_i = R_j + \sigma_j z \sqrt{1 - r^2} + r\sigma_j A$, where ρ_i was the stochastic rate, z is a z -value (see below), r was the correlation coefficient between this rate and 3+ female mortality, σ_j^2 is the standard deviation of the vital rate and A was the average value of the z values for 3+ female mortality rates. The values of r were -0.452 for fecundity and 0.522 for male mortality. The z values for population 1 were simple standardized normal deviates. Those for population 2 were calculated as: $z_2 = Z_2 \sqrt{1 - r^2} + rz_1$, where Z_2 was the original z value, r was the correlation between populations, and z_1 was the z -value calculated for the equivalent age class and vital rate for population 1. The value of r was 0.7 .

Received 2 August; accepted 20 November 2001.

- Klein, D. R. The introduction, increase, and crash of reindeer on St. Matthew Island. *J. Wildl. Mgmt* **32**, 350–367 (1968).
- Flood, D. R. A study of sex differential mortality in the survival of wapiti. (Canadian Wildlife Service Report Series, Ottawa, 1970).
- Clutton-Brock, T. H., Price, O. F., Albon, S. D. & Jewell, P. A. Persistent instability and population regulation in Soay sheep. *J. Anim. Ecol.* **60**, 593–608 (1991).
- Clutton-Brock, T. H., Guinness, F. E. & Albon, S. D. *Red Deer: Behaviour and Ecology of Two Sexes* (Univ. Edinburgh Press, Edinburgh, 1982).
- Clutton-Brock, T. H. & Albon, S. D. *Red Deer in the Highlands* (Blackwell Scientific, Oxford, 1989).
- Clutton-Brock, T. H., Major, M. & Guinness, F. E. Population regulation in male and female red deer. *J. Anim. Ecol.* **54**, 831–846 (1985).
- Clutton-Brock, T. H. & Loneragan, M. E. Culling regimes and sex ratio biases in Highland red deer. *J. Appl. Ecol.* **31**, 521–527 (1994).
- Bryden, J. M. in *The Next Twenty Years* 1–41 (Red Deer Commission, Inverness, 1979).
- Jarvis, E. *The Red Deer Industry* (The Scottish Landowner's Federation, Edinburgh, 1978/9).
- Milne, J. A., Birch, C. P. D., Hester, A. J., Armstrong, H. & Robinson, A. The impact of vertebrate herbivores on the natural vegetation of the Scottish upland: a review. 1–127 (Report no. 95, Scottish National Heritage, Edinburgh 1998).
- Lowe, V. P. W. Population dynamics of the red deer (*Cervus elaphus* L.) on Rhum. *J. Anim. Ecol.* **38**, 425–457 (1969).
- Lowe, V. P. W. in *The Scientific Management of Animal and Plant Communities for Conservation* (eds Duffey, E. & Watt, A. S.) 437–445 (Blackwell, Oxford, 1971).
- Kruuk, L. E. B., Clutton-Brock, T. H., Albon, S. D., Pemberton, J. M. & Guinness, F. E. Population density affects sex ratio variation in red deer. *Nature* **399**, 459–461 (1999).
- Clutton-Brock, T. H., Albon, S. D. & Guinness, F. E. Parental investment and sex differences in juvenile mortality in birds and mammals. *Nature* **313**, 131–133 (1985).
- Clutton-Brock, T. H., Rose, K. E. & Guinness, F. E. Density-related changes in sexual selection in red deer. *Proc. R. Soc. Lond. B* **264**, 1509–1516 (1997).
- Charles, W. N., McCowan, D. & East, K. Selection of upland swards by red deer *Cervus elaphus* L. on Rhum. *J. Appl. Ecol.* **14**, 55–64 (1977).
- Clutton-Brock, T. H., Iason, G. R. & Guinness, F. E. Sexual segregation and density-related changes in habitat use in male and female red deer (*Cervus elaphus*). *J. Zool.* **211**, 275–289 (1987).
- Conradt, L., Clutton-Brock, T. H. & Thomson, D. Habitat segregation in ungulates: are males forced into suboptimal habitats through indirect competition by females? *Oecologia* **119**, 367–377 (1999).
- Mutch, W. E. S., Lockie, J. D. & Cooper, A. N. *The Red Deer in South Ross: a Report on Wildlife Management in the Scottish Highlands* (Department of Forestry and Natural Resources, Univ. Edinburgh, 1976).
- Annual Report 2000–2001. (Deer Commission for Scotland, Inverness, 2001).
- Langvatn, R. & Loison, A. Consequences of harvesting on age structure, sex ratio and population dynamics of red deer *Cervus elaphus* in central Norway. *Wildl. Biol.* **5**, 213–223 (1999).
- Mitchell, B., Staines, B. W. & Welch, D. *Ecology of Red Deer: a Research Review Relevant to their Management in Scotland* (Institute of Terrestrial Ecology, Cambridge, 1977).
- Conradt, L., Clutton-Brock, T. H. & Guinness, F. E. Sex differences in weather sensitivity as cause of habitat segregation in red deer (*Cervus elaphus* L.). *Anim. Behav.* **59**, 1049–1060 (2000).
- Osborne, B. C. Habitat use by red deer (*Cervus elaphus*) and hill sheep in the west Highlands. *J. Appl. Ecol.* **21**, 497–506 (1984).
- Clutton-Brock, T. H. & Albon, S. D. Trial and error in the Highlands. *Nature* **358**, 11–12 (1992).
- Sibbald, A. Using GPS to track wild red deer stags. *Deer* **11**, 524–529 (2000).
- Buckland, S. T., Ahmadi, S., Staines, B. W., Gordon, I. J. & Youngson, R. W. Estimating the minimum population size that allows a given annual number of mature red deer stags to be culled sustainably. *J. Appl. Ecol.* **33**, 118–130 (1996).
- Holloway, C. W. *The Effect of Red Deer and Other Animals on Naturally Regenerated Scots Pine*. Thesis, Univ. Aberdeen (1967).
- McCullough, D. R., Pire, D. S., Whitmore, D. L., Mansfield, T. M. & Decker, R. H. Linked sex harvest strategy for big game management with a test case in black-tailed deer. *Wildl. Monog.* **112**, 1–41 (1990).
- McCullough, D. R. Population manipulations of North American deer *Odocoileus* spp.: balancing high yield with sustainability. *Wildl. Biol.* **7**, 161–171 (2001).
- Lubow, B. C., White, G. C. & Anderson, D. R. *J. Wildl. Mgmt* **60**, 787–796 (1996).

Acknowledgements

We thank F. E. Guinness, J. M. Pemberton and S. Albon for maintaining the long-term records in the north block of Rum; C. Duck, J. Kingsley, M. Twiss and C. Covey for collecting data on deer populations throughout Rum; R. Scott for organizing the manipulations of deer numbers on Rum; the staff of the Deer Commission for Scotland, especially C. MacLean and D. Youngson, for counting and culling deer on Rum; the SNH staff on Rum, especially M. Curry and D. Reed, for support in maintaining deer research on the island; S. D. Albon, I. Gordon, L. Conradt, L. Kruuk and S. Buckland for comments or help with these analyses; J. Taylor for providing Fig. 1b; and the NERC, the Deer Commission for Scotland, SNH and the Scottish Executive for funding.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.H.C.-B. (e-mail: thcb@hermes.cam.ac.uk).

Influence of scene statistics on colour constancy

Jürgen Golz* & Donald I. A. MacLeod†

* Department of Psychology, University of Kiel, 24098 Kiel, Germany

† Department of Psychology, University of California at San Diego, La Jolla, California 93093-0109, USA

The light reflected from an object depends not only on the surface properties of this object but also on the illuminant. The same is true for the excitations of the photoreceptors, which serve as the basis for the perceived colour. However, our visual system has the ability to perceive constant surface colours despite changes in illumination¹. The average chromaticity of the retinal image of a scene depends on the illumination, and thus might be used by the visual system to estimate the illumination and to modulate the correction that subserves colour constancy^{2–4}. But this measure is not sufficient: a reddish scene under white light can produce the same mean stimulation as a neutral scene in red light. Higher order scene statistics—for example, the correlation between redness and luminance within the image—allow these cases to be distinguished. Here we report that the human visual system does exploit such a statistic when estimating the illuminant, and gives it a weight that is statistically appropriate for the natural environment.

A viewed surface reflects towards the observer's eye a characteristic fraction of the light that is cast on it; therefore, the stimulus for vision is the product of an illuminant component and a surface-reflectance component. Colour constancy requires the visual system to disentangle the surface-reflectance component from the product—in some sense discounting the illuminant colour—but the illuminant colour is not generally directly known. Problems like this are common in many aspects of visual perception. They are mathematically underdetermined, and additional constraints must be drawn from the regularities of the physical world to achieve a unique solution.

The mean chromaticity of the retinal image is often proposed to be an indicator of illuminant colour^{2–4}, because for a given scene this statistic varies systematically with changes in illumination. But in more realistic situations in which the scenes might change as well, this measure is ambiguous (Fig. 1a). To illustrate the ambiguity, consider the problem of distinguishing a predominantly reddish room under white light from a predominantly neutral one under reddish light. The mean chromaticity of the image received by the eye may be the same in these two cases, but human observers can

resolve the ambiguity and perceive correctly the reddish room as redder^{1,5}. Even in scenes where obvious cues to the illuminant contained in the spatial structure of the image, such as specular reflections^{6–9} and mutual reflections^{10–12}, are not available, the visual system can distinguish changes of the mean image chromaticity caused by illuminant changes from changes of the mean chromaticity caused by variations of the set of surfaces¹³.

Here we offer an explanation for this ability: in such situations the distribution of intensity and chromaticity for the various surfaces

within the retinal image yields additional evidence about the illuminant colour, and this evidence is exploited in human colour perception. In particular, as we show below, the interplay of surfaces and illuminants causes the correlation between surface redness and luminance within an image to be useful for resolving the image into its illuminant and surface components. Two examples of how this higher order statistic might resolve the ambiguity of the mean image chromaticity and thus improve the estimation of the illumination are shown in Fig. 1b and c.

We tested experimentally whether the luminance–redness correlation is used by the human visual system for estimating the illuminant colour. We analysed images of natural scenes¹⁴ to find out how effective this and other higher order statistics are as information about the chromaticity of surfaces and illuminants in the natural environment.

For our experiments, we adopted stimuli that make it possible to vary independently various statistics (means, variances, correlations) of the distribution of colour and lightness within the display^{15–17}. A circular test field was surrounded by random patterns of overlapping circles; these were of a fixed diameter but varied in colour and luminance to a degree that is typical of natural scenes¹⁸. We asked subjects to adjust the colour of the test field so that it appeared neutral grey. We chose a logarithmic cone excitation space (with the axes $\log r$, $\log b$ and $\log$ luminance) as our coordinate system. The chromaticity value r is the luminance-normalized excitation of cones sensitive to long wavelength, and b is the luminance-normalized excitation of cones sensitive to short wavelength¹⁹. An equal energy white has the values $r = 0.7$ and $b = 1.0$ ($\log r = -0.155$ and $\log b = 0.0$). Reds have higher r values, and greens have lower ones; blues have higher b values, and yellows have lower ones. Our experiments and analysis focused on the r coordinate, which varied over a range from distinctly greenish to distinctly reddish.

In our main experiment^{14,20}, we varied the luminance–redness correlation for the elements surrounding the test field—by introducing a linear dependence between $\log r$ and $\log$ luminance— independently of other statistics (means and variances). For a given condition, the chromaticity and luminance values for the circles in the surround were chosen to achieve a certain correlation value (-1.0 , -0.8 , 0.0 , 0.8 or 1.0). If the perceived colour of the centre test spot was not influenced by the varied correlation, then the settings to make the test spot neutral grey would be the same for all five conditions, because the space-averaged chromaticities of the surrounds were the same. The surrounds would then be functionally equivalent with respect to the perceived colour of the centre test spot.

The results for subject DB are shown in Fig. 2a. For conditions

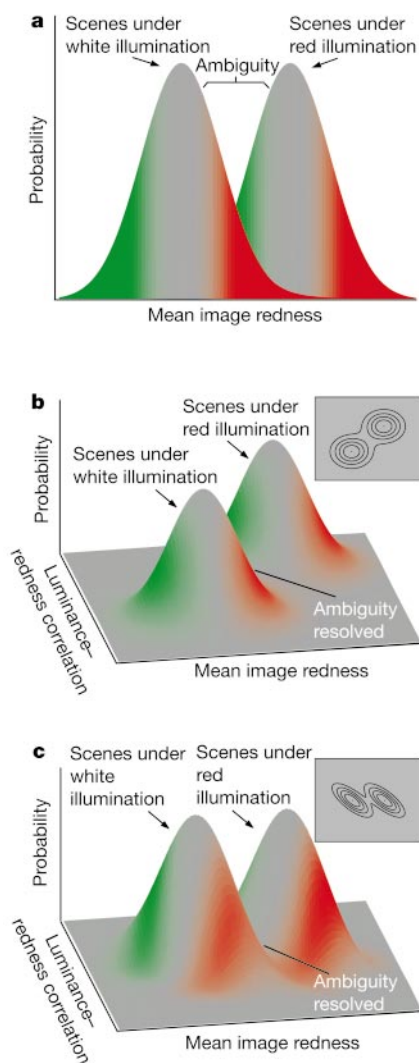


Figure 1 The ambiguity of the mean image chromaticity for estimating the illuminant, and two possibilities for resolving it using a higher order statistic (luminance–redness correlation). **a**, The same mean image chromaticity could result either from a reddish scene under white illumination or from a white scene under reddish illumination, because these two illuminants, applied to the given sample of scenes, generate overlapping distributions of mean image redness. **b**, Taking into account higher order statistics might resolve this ambiguity. In this example a higher order statistic, the luminance–redness correlation over the pixels of a single image, is increased for all images under the reddish illumination. Thus, a high luminance–redness correlation within the image would be evidence that the illuminant is reddish. Inset shows the corresponding view from above, depicting probability values by contour lines. **c**, Another way in which higher order scene statistics might separate the distributions of images from two different illuminants and thus reduce the ambiguity encountered in considering mean chromaticity alone. By evaluating both the mean and correlation of a particular image, an observer could estimate two unknowns: the degree of redness inherent in the objects making up the scene; and the redness of the light source that illuminates the scene. Inset shows the corresponding view from above, depicting probability values by contour lines.

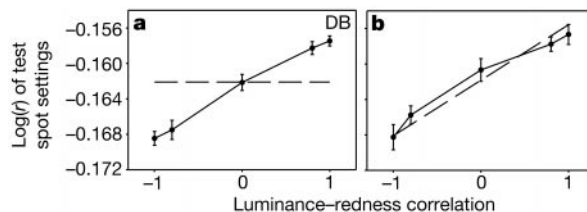


Figure 2 Dependence of centre test spot settings on the luminance–redness correlation within the surround. **a**, Results for subject DB. Filled circles are the mean $\log r$ values for perceptually achromatic test fields plotted as a function of the correlation between $\log$ luminance and $\log r$ for the five surround conditions; error bars represent ± 1 s.e.m. If the settings were not dependent on the correlation but only on the mean of the chromaticity of the surround, then settings would be the same for all conditions (horizontal dashed line). **b**, Average results for all subjects (circles) compared with predictions based on optimal use of the correlation statistic (dashed line). Error bars for the experimental results represent $\pm$ s.e.m. for subject variability.

with higher correlation between redness and luminance, a more reddish chromaticity was required to make the test field subjectively achromatic. The data for eight out of ten subjects tested were quantitatively similar and individually statistically significant ($P < 0.001$, linear trend test; two of these eight observers were also tested for correlation values of -0.4 and $+0.4$ and the results were in proportion to the other values). When the correlation between redness and luminance was positive, subjects selected a physically more reddish (higher $\log r$) test field as neutral grey. Because higher $\log r$ values are associated with redder illumination of a physically neutral surface, this is the result that is expected if the observer infers a more reddish illumination in the case of a positive luminance–redness correlation and perceives neutral grey when a correspondingly reddish light stimulus is received from the test field. As shown in Fig. 2, this effect of the luminance–redness correlation is far from trivial; by comparison, a just noticeable difference in $\log r$ is only about 0.0004 (for example, see ref. 21).

Other work has provided evidence that the correction that subserves colour constancy is not governed merely by the space-averaged chromaticity^{1,13,15–17,22–24}, and some current colour constancy algorithms provide general frameworks for effects of this sort^{25–28}. But so far only one group has made a specific suggestion about what statistic might also be diagnostic of the illuminant and found evidence that a reduced chromatic variance serves as a cue for a ‘non-normal’ illumination^{15–17}. The luminance–redness correlation that we suggest makes it possible to resolve the retinal image into its illuminant and surface components over the whole range of typical daylight, as the following analysis shows.

To check how well chromatic statistics in images of natural scenes can support inferences about illuminant and surface colours, we used hyperspectral images of natural scenes¹⁸. To these 12 scenes, 4 daylight illuminations²⁹ were applied (correlated colour temperatures 4,000 K, 5,500 K, 8,500 K and 20,000 K; with increasing correlated colour temperature, the illuminants become less reddish (lower r values) and more bluish (higher b values)). We calculated

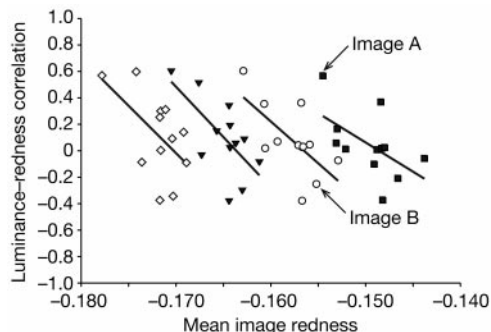


Figure 3 Luminance–redness correlation and mean redness of natural scenes under different illuminations. The vertical coordinate of each data point represents the correlation between pixel redness, $\log r$ and pixel luminance, $\log$ luminance, within an image of a scene under a particular illuminant. This correlation is plotted against mean image redness (mean of $\log r$). The four clusters show data for 12 natural scenes under four different illuminants (filled squares, colour temperature 4,000 K; open circles, 5,500 K; filled triangles, 8,500 K; open diamonds, 20,000 K). The rank order for mean image redness of the 12 scenes stays almost unchanged under all four illuminants, indicating that scene redness can be roughly assessed by mean image redness for a given illuminant. For a given scene, the luminance–redness correlations are almost independent of illumination, but within each illuminant cluster they are more negative for the redder scenes, as shown by the negatively sloped regression lines. Thus, correlation and mean together separate the distributions of images resulting from different illuminants and make it possible to distinguish scene redness from illuminant redness, much as in the case of Fig. 1c. (The correlation between correlation ($\log r$, $\log$ luminance) and mean ($\log r$) for the four illuminant clusters are -0.555 (filled squares), -0.654 (open circles), -0.653 (filled triangles), -0.556 (open diamonds), in each case significantly different from zero, $P < 0.05$).

several statistics (mean, variance, correlation, skewness) of the luminance and chromaticities for each illuminated scene. Of all the statistics that we evaluated besides the means, only the luminance–redness correlation was found to be useful for estimating the illumination colour.

Although the experimental results suggest that the human visual system interprets a high luminance–redness correlation as a redder illuminant, the correlation alone is not an effective cue in determining the illuminant colour (Fig. 3). The correlations are almost independent of illumination. This measure can nevertheless resolve ambiguity, but more like in the case shown in Fig. 1c than in Fig. 1b; that is, when used together with the mean image redness, this measure separates the clusters of images from different illuminations (the four diagonally oriented clusters in Fig. 3) and thus permits the estimation of illumination colour.

For example, an image with a mean $\log r$ value of -0.155 could be a reddish scene under neutral light or a neutral (or even greenish) scene under reddish light (Fig. 3). But if the visual system takes into account the correlation between redness and luminance, it can distinguish between illumination redness and scene redness. The more positive the correlation, the greater the evidence for the redness of the illumination as opposed to redness of the scene. In Fig. 3, image A is with high probability a greenish scene under the reddish illuminant of 4,000 K whereas image B is likely to be a reddish scene that belongs to the cluster of the more neutral illuminant of 5,500 K.

The reason why the luminance–redness correlation can help in this way is that scene redness and illuminant redness affect the correlation differently. Because of the eye’s spectral sensitivity, middle wavelengths contribute more to luminance than do short or long wavelengths. Under ‘neutral’ lighting, it is the neutral or moderately reddish and greenish pixels that reflect most middle-wavelength light, and which therefore have the highest luminance. In neutral or greenish scenes, surfaces that reflect predominantly long wavelengths and thus have lower luminance are rare. But in reddish scenes under neutral lighting, the eye’s diminishing sensitivity at long wavelengths will discriminate against the most reddish pixels of a reddish scene, and these will therefore tend to be of much lower luminance. The luminance–redness correlation thus becomes more negative with more reddish scenes, as the sloped regression lines in Fig. 3 show.

No comparable effect occurs if it is not the scene but the light that is reddish. In this case, the low luminosity of reds is counterbalanced by the illuminant’s greater energy at long wavelengths. Reddish surfaces reflect a larger proportion of long-wavelength than of short-wavelength light. Thus, when the illuminant becomes reddish the reddish surfaces will be more luminous relative to other colours in the scene, owing to the better overlap of their spectral reflectances with the spectral range in which the illuminant has its highest power.

Thus, reddish scenes but not reddish illuminants generate images with negative luminance–redness correlations. By evaluating both mean and correlation—two independent quantities—an observer can estimate two unknowns: the predominant colour (in this case, the degree of redness) inherent in the objects making up the scene; and the redness of the light source that illuminates the scene. In this way, the ambiguity encountered in considering mean chromaticity alone can be resolved.

How much weight should a smart visual system give to the correlation between redness and luminance in estimating the illumination? To answer this question for our simulated world of natural scenes under different illuminations, we calculated a maximum likelihood estimate for the chromaticity of the illumination given the mean and correlation value of an image (see Methods). Figure 2b illustrates the results in terms of how an optimal observer, adopting this maximum likelihood estimate, would have reacted to the stimuli of our experiment. The dashed line is a parameter-free

prediction for the test field chromaticity that would appear neutral grey, which is based on the hypothesis that optimal weight is given to the correlation measure in estimating the illuminant. For comparison, the circles in Fig. 2b show the test field settings that appeared neutral by experiment. The size of the observed effect of this statistic on colour perception is roughly consistent with optimal computation. □

Methods

Experiments

Stimuli consisted of a random pattern of overlapping circles (each 3.6° in diameter) covering the whole display area of 41° (height) and 53° (width). The chromaticity distributions of these displays, set by assigning different colours to the circles, were -0.16 ± 0.01 and -0.14 ± 0.06 (mean $\pm$ s.d.; $\log r$ and $\log b$). Luminance was 7 ± 3.6 cd m $^{-2}$ (mean $\pm$ s.d.). The statistics of the correlation between $\log r$ and $\log b$ and luminance were -1.0 , -0.8 , 0.0 , 0.8 and 1.0 for the five conditions, whereas the correlation between $\log b$ and $\log r$ was always 0.0 , and the correlation between $\log b$ and $\log r$ was always -0.137 . The test spot was a circle in the centre of the display, which was also 3.6° in diameter but on top of all other circles in this area. The luminance of the test spot was held constant at 7 cd m $^{-2}$ while the subjects adjusted its chromaticity in the ($\log r$, $\log b$) plane by means of a trackball to make it look grey.

We presented stimuli on a carefully calibrated 17-inch (43-cm) Mitsubishi Diamond Pro monitor driven by a CRS VSG2/4f graphic board with 15 bits resolution per gun. Subject viewed the stimuli from a distance of 55 cm in a dark room through a tunnel of black velvet to exclude any other light sources or reflections. To achieve an equilibrium state of adaptation, subjects had to adapt to each display for at least 100 s.

Ten subjects aged between 11 and 34 years participated in two 1-h sessions making six settings per condition. All participants except one (J.G.) were naive to the purpose of the experiment. All had normal or corrected-to-normal vision and showed no colour vision deficiency, as tested by Ishihara's test plates.

Theoretical analysis

The maximum likelihood estimate of the illumination of an image of a natural scene given its mean redness and luminance—redness correlation was determined by means of a linear regression. We adopted as predictors X_1 , the mean of the decimal log of r , and X_2 , the correlation between log luminance and $\log r$. These predictors could be specified for each of the 12 natural scenes under each of the four illuminants. The optimal weights derived for the prediction of the $\log r$ value of the illuminant ($\hat{Y}$) are embodied in the following linear equation: $\hat{Y} = 1.1305X_1 + 0.0063X_2 + 0.0202$. In deriving the theoretical straight line of Fig. 2b, we assumed that an optimal observer, when confronted with our experimental stimuli, would estimate the redness of the illumination according to this equation in each case and would perceive the resulting estimated illuminant chromaticity as neutral grey, effectively discounting the estimated illuminant in the perception of physically neutral surfaces.

We used the Stockman–MacLeod–Johnson cone sensitivities³⁰ to calculate all cone excitation values.

Received 27 March; accepted 27 November 2001.

- Kraft, J. M. & Brainard, D. H. Mechanisms of color constancy under nearly natural viewing. *Proc. Natl Acad. Sci. USA* **96**, 307–312 (1999).
- Buchsbaum, G. A. Spatial processor model for object colour perception. *J. Franklin Inst.* **310**, 1–26 (1980).
- Land, E. H. Recent advances in retinex theory. *Vis. Res.* **26**, 7–21 (1986).
- Gershon, R. & Jepson, A. D. The computation of color constant descriptors in chromatic images. *Color Res. Appl.* **14**, 325–334 (1989).
- Gilchrist, A. L. & Ramachandran, V. Red rooms in white light appear different from white rooms in red light. *Invest. Ophthalmol. Vis. Sci.* **33**, 756 (1992).
- Lee, H.-C. Method for computing the scene–illuminant chromaticity from specular highlights. *J. Opt. Soc. Am. A* **3**, 1694–1699 (1986).
- Tominaga, S. & Wandell, B. A. Standard surface-reflectance model and illuminant estimation. *J. Opt. Soc. Am. A* **6**, 576–584 (1989).
- D'Zmura, M. & Lennie, P. Mechanisms of color constancy. *J. Opt. Soc. Am. A* **3**, 1662–1672 (1986).
- Maloney, T. M. & Yang, J. N. in *Colour Perception: From Light to Object* (eds Mausfeld, R. & Heyer, D.) (in the press).
- Funt, B. V. & Drew, M. S. Color space analysis of mutual illumination. *IEEE Trans. Pattern Anal. Mach. Intell.* **15**, 1319–1326 (1993).
- Funt, B. V., Drew, M. S. & Ho, J. Color constancy from mutual reflection. *Int. J. Comp. Vis.* **6**, 5–24 (1991).
- Bloj, M., Kersten, D. & Hurlbert, A. C. Perception of three-dimensional shape influences colour perception through mutual illumination. *Nature* **402**, 877–879 (1999).
- Bäuml, K.-H. Color appearance: effects of illuminant changes under different surface collections. *J. Opt. Soc. Am. A* **11**, 531–542 (1994).
- MacLeod, D. I. A., & Golz, J. in *Colour Perception: From Light to Object* (eds Mausfeld, R. & Heyer, D.) (in the press).
- Mausfeld, R. Color in *Color Vision* (eds Backhaus, W. G. K., Kliegel, R. & Werner, J. S.) 219–250 (De Gruyter, Berlin, 1998).
- Mausfeld, R. & Andres, J. A reduced variance of receptor codes in chromatic scenes activates a 'discounting the illuminant' mechanism. *Perception* **27** (Suppl.), 42 (1998).

- Mausfeld, R. & Andres, J. Second order statistics of colour codes modulate transformations that effectuate varying degrees of scene invariance and illumination invariance. *Perception* (in the press).
- Ruderman, D. L., Cronin, T. W. & Chiao, C. C. Statistics of cone responses to natural images: implications for visual coding. *J. Opt. Soc. Am. A* **15**, 2036–2045 (1998).
- MacLeod, D. I. A. & Boynton, R. M. Chromaticity diagram showing cone excitation by stimuli of equal luminance. *J. Opt. Soc. Am.* **69**, 1183–1186 (1979).
- Golz, J. & MacLeod, D. I. A. Influence of scene statistics on color constancy. *Invest. Ophthalmol. Vis. Sci.* **40** (Suppl.), S749 (1999).
- Eskew, R. T., McLellan, J. S. & Giulianini, F. in *Color Vision: from Genes to Perception* (eds Gegenfurtner, K. R. & Sharpe, L. T.) 345–368 (Cambridge Univ. Press, New York, 1999).
- Jenness, J. W. & Shevell, S. K. Color appearance with sparse chromatic context. *Vis. Res.* **35**, 797–805 (1995).
- Brown, R. O. & MacLeod, D. I. A. Color appearance depends on the variance of surround colors. *Curr. Biol.* **7**, 844–849 (1997).
- Webster, M. A. & Mollon, J. D. Adaptation and the color statistics of natural images. *Vis. Res.* **37**, 3283–3298 (1997).
- Brainard, D. H. & Freeman, W. T. Bayesian color constancy. *J. Opt. Soc. Am. A*, **14**, 1393–1411 (1997).
- D'Zmura, M., Iverson, G. & Singer, B. in *Geometric Representations of Perceptual Phenomena* (eds Luce, D., D'Zmura, M., Hoffman, D., Iverson, G. & Romney, A.) 187–202 (Lawrence Erlbaum Associates, Mahwah, 1995).
- Forsyth, D. A. in *AI and the Eye* (ed. Blake, A. & Troscianko, T.) 201–227 (Wiley, Chichester, 1990).
- Forsyth, D. A. A novel algorithm for color constancy. *Int. J. Comp. Vis.* **5**, 5–36 (1990).
- Wyszecki, G. & Stiles, W. S. *Color Science: Concepts and Methods, Quantitative Data and Formulae* 2nd edn 145–146 (Wiley, New York, 1982).
- Stockman, A., MacLeod, D. I. A. & Johnson, N. E. Spectral sensitivities of the human cones. *J. Opt. Soc. Am. A* **10**, 2491–2521 (1993).

Acknowledgements

We thank D. L. Ruderman, T. W. Cronin and C. C. Chiao for their spectral data of natural scenes. This work was supported by the National Eye Institute. J. Golz was supported by the German–American Fulbright Commission.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.G. (e-mail: golz@psychologie.uni-kiel.de).

Early consolidation in human primary motor cortex

Wolf Muellbacher*†, Ulf Ziemann*†, Joerg Wissel‡, Nguyet Dang*, Markus Kofler‡, Stefano Facchini*, Babak Boroojerdi*, Werner Poewe‡ & Mark Hallett*

* *Human Motor Control Section, Medical Neurology Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bldg 10, Rm 5N226, 10 Center Drive MSC 1428, Bethesda, Maryland 20892-1428, USA*
 ‡ *Neurological Department, University Hospital of Innsbruck, Anichstrasse 35, A-6020 Innsbruck, Austria*

Behavioural studies indicate that a newly acquired motor skill is rapidly consolidated from an initially unstable state to a more stable state¹, whereas neuroimaging studies demonstrate that the brain engages new regions for performance of the task as a result of this consolidation². However, it is not known where a new skill is retained and processed before it is firmly consolidated. Some early aspects of motor skill acquisition involve the primary motor cortex (M1)³, but the nature of that involvement is unclear. We tested the possibility that the human M1 is essential to early motor consolidation. We monitored changes in elementary motor behaviour while subjects practised fast finger movements that rapidly improved in movement acceleration and muscle force

† Present addresses: Ludwig Boltzmann Institute for Epilepsy and Neuromuscular Disorders, Neurological Hospital of Vienna, Rosenhugel, Riedelgasse 5, A-1130, Vienna, Austria (W.M.); Clinic of Neurology, JW Goethe University of Frankfurt, Theodor-Stern-Kai 7, D-60590 Frankfurt-am-Main, Germany (U.Z.).

generation. Here we show that low-frequency, repetitive transcranial magnetic stimulation of M1 but not other brain areas specifically disrupted the retention of the behavioural improvement, but did not affect basal motor behaviour, task performance, motor learning by subsequent practice, or recall of the newly acquired motor skill. These findings indicate that the human M1 is specifically engaged during the early stage of motor consolidation.

Only recently has it been determined that motor skills are consolidated as part of the learning process¹. Consolidation of a new skill is disrupted when a second skill is learned immediately after the first. There is no disruption if 4 h elapse between learning the two motor skills, with consolidation occurring gradually over this period, indicating a time-dependent process. Neuroimaging has shown that when a new motor skill is consolidated, there is a shift of brain activity from prefrontal regions to the premotor, posterior parietal and cerebellar cortex structures². However, the first steps in motor consolidation are unclear, although several studies have indicated that the primary motor cortex (M1) is involved in motor learning in some way^{3,4}. In a series of experiments, we have shown that the human M1 is involved early in motor learning processes⁵. Using repetitive transcranial magnetic stimulation (rTMS), a procedure that can interfere with local cortical brain function, we demonstrate here in six separate experiments that the human M1 is crucial to the early stage of motor memory consolidation.

First, we studied learning effects of brief motor practice (MP) on elementary motor behaviour. Changes in peak pinch force and peak pinch acceleration were monitored while naive subjects practised a ballistic pinch of the index finger and thumb to the 0.5-Hz beat of a metronome. Each of two 5-min practice episodes (practices 1 and 2) was followed by a 15-min rest, which was followed by another 10-min

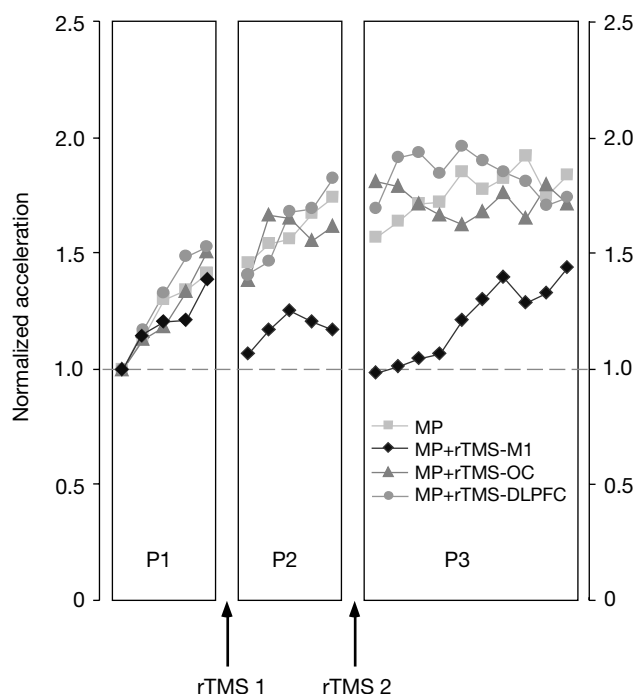


Figure 1 Effects of motor practice and rTMS on motor learning and early motor consolidation: acceleration data. Each symbol represents the (normalized) mean peak acceleration for each practice condition. During practice 1 (P1) there was an increase in peak pinch acceleration under all conditions. rTMS of M1 cancelled the retention of the behavioural improvement of practices 1 and 2 (P2). The ability to improve behaviour by subsequent practice (P3) was unimpaired, but the final improvement was less marked than in the subjects who did not receive rTMS of M1. rTMS of the occipital cortex or dorsolateral prefrontal cortex had no impact on early motor consolidation.

practice (practice 3). Subjects rapidly learned to optimize the ballistic finger movements, as indicated by an increasing peak pinch acceleration during practice 1 (correlation coefficient $r = 0.97$, $P = 0.005$). They showed both retention of this behavioural improvement and additional improvement during further practising, resulting in significant gains in acceleration ($r = 0.91$, $P < 0.0001$) and force ($P = 0.003$) at the end of the 20-min learning period (Figs 1, 2). This learning-related improvement in elementary motor behaviour is associated with task- and effector-specific changes in M1 excitability⁵, consistent with previous reports of learning-related changes of M1^{3,6–8}.

Second, we explored the role of M1 in early motor consolidation. rTMS of M1 was applied immediately after practices 1 and 2 to interfere with M1 function, in an attempt to disrupt early motor consolidation (MP+rTMS-M1). rTMS can depress local cortical function that outlasts the stimulation duration^{9–11}. The retention of the significant increase in peak acceleration gained during practice 1 was lost after rTMS of M1 ($P = 0.037$; Fig. 1). During practice 2, peak acceleration again increased, but, for a second time, rTMS cancelled the retention of the behavioural improvement. Peak acceleration once more increased during practice 3, but the total behavioural gain at the end of the practice period was less than in the subjects who did not receive rTMS of M1 (Fig. 1). Strategy, number of successful/unsuccessful movements ($P = 0.389$) and error (incorrect or no movement) rate ($P = 0.107$) were not different from the MP experiment, indicating that the rTMS effects could not simply be explained by a nonspecific change in motor performance. Thus, rTMS of M1 cancelled the behavioural gain, whereas motor performance and the ability to learn during subsequent practice were not impaired. This indicates that rTMS of M1 had disrupted early motor consolidation.

The specificity of these rTMS effects on early motor consolidation was demonstrated in four control experiments. First, we explored whether the effects of rTMS were caused by interference with basal motor behaviour. For this, maximal peak force and peak acceleration were measured immediately before and after rTMS of M1 in a group of subjects who did not practice (rTMS-M1). Both peak force ($P = 0.679$; Fig. 2) and peak acceleration (mean $\pm$ s.d. 1.27 ± 0.533 ,

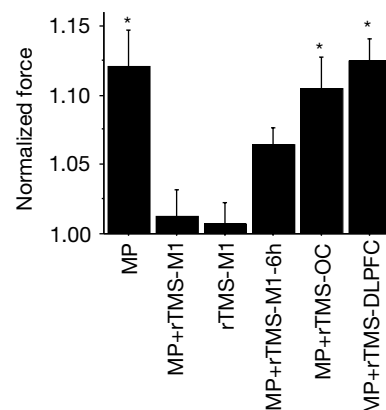


Figure 2 Effects of motor practice and rTMS on motor learning and early motor consolidation: force data. Each bar represents the mean peak pinch force at the end of each experiment expressed relative (as a quotient) to the mean peak pinch force obtained from the beginning of the respective experiment. Motor practice (MP) resulted in an increase in peak pinch force. In contrast, there was no such increase in the MP+rTMS-M1 group. rTMS of M1 did not affect basal peak force *per se*. Neither rTMS of the occipital cortex (MP+rTMS-OC) nor of the dorsolateral prefrontal cortex (MP+rTMS-DLPFC) interfered with the behavioural improvement (the 20-min practice effects were similar in the MP, MP+rTMS-OC and MP+rTMS-DLPFC groups). There was a slightly lesser increase in peak pinch force in the MP+rTMS-M1-6h group, which can be explained by the shorter practice duration of 10 min in this condition compared with 20 min in the other practice conditions. Error bars are s.e.m. Asterisk indicates $P < 0.05$ (paired *t*-test).

range 0.55–2.22; $P = 0.187$) were unaffected by the stimulation procedure, demonstrating that rTMS of M1 does not impair basal motor behaviour *per se*.

Second, we considered whether cancellation of the retention of the behavioural improvement could be due to interference with the recall of the newly acquired motor skill. For this, subjects practised for 5 min (practice 1), and rTMS of M1 was applied after a 6-h consolidation period¹ (MP+rTMS-M1-6h). The failure of rTMS to interfere with the improved behaviour after these 6 h ($P = 0.86$; Fig. 3) shows that the new motor skill had become resistant to interruption. This is a crucial finding to conclude that not the recall of the newly learned motor skill but specifically the early motor consolidation was disrupted in the MP+rTMS-M1 experiment. Furthermore, the MP+rTMS-M1-6h experiment excludes the possibility that rTMS of M1 is sufficient to nonspecifically disrupt the retention of the new motor skill.

Lastly, we controlled whether early motor consolidation is also disrupted if rTMS is applied to the occipital cortex (MP+rTMS-OC) or to the dorsolateral prefrontal cortex (MP+rTMS-DLPFC), two brain cortex structures reported to be involved in early motor learning processes². The failure to disrupt the behavioural improvement by rTMS of these cortex structures (acceleration: MP+rTMS-OC, $P = 0.135$; MP+rTMS-DLPFC, $P = 0.229$; force: MP+rTMS-OC, $P = 0.186$; MP+rTMS-DLPFC, $P = 0.299$) (Figs 1, 2) is evidence that the occipital cortex and the dorsolateral prefrontal cortex are not crucial to early motor consolidation.

We wondered what neural elements were disturbed by rTMS over M1. We used TMS-induced activation of small hand muscles to determine the optimal scalp position for rTMS so that the target hand M1 was definitively in the centre of the stimulation field. To test whether rTMS at this site actually interfered with M1, we studied rTMS-induced changes in single-pulse TMS measures of M1 excitability. Motor threshold and amplitude of motor-evoked potential (MEP) were used as measures of M1 excitability before and after the rTMS procedure. The demonstration that rTMS over hand M1 increases motor threshold (MP+rTMS-M1, rTMS-M1

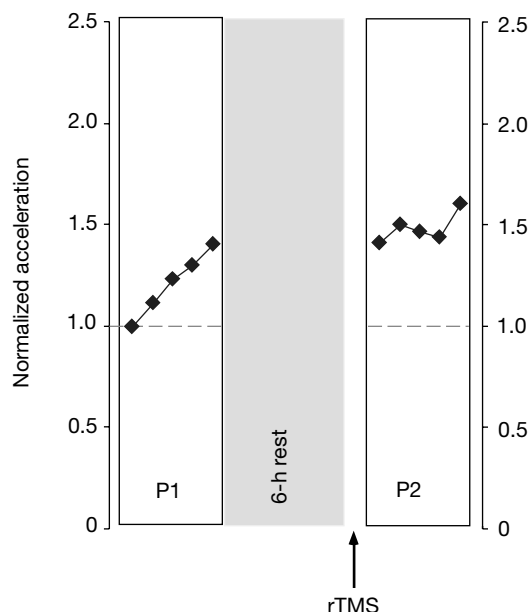


Figure 3 Effects of rTMS of M1 on the recall of the newly learned motor task: MP+rTMS-M1-6h experiment. This graph illustrates the effects of rTMS of M1 when applied 6 h after the practice (each symbol represents the normalized mean peak acceleration). During practice 1 (P1) there was an increase in peak pinch acceleration, which was similar to the other practice experiments (see Fig. 1). rTMS of M1 after a 6-h consolidation period did not disrupt the recall of the newly acquired motor skill.

and MP+rTMS-M1-6h experiments; $P < 0.0001$; Fig. 4a) and depresses MEP amplitude (rTMS-M1 and MP+rTMS-M1-6h experiments; $P < 0.005$; Fig. 4b) is proof that rTMS over M1 significantly interfered with M1. The failure of rTMS to depress MEP amplitude if subjects performed motor practice immediately before rTMS (MP+rTMS-M1 rather than the rTMS-M1 and MP+rTMS-M1-6h experiments; Fig. 4b) indicates post-practice activity in M1 during the early motor consolidation period. This post-practice activity (increased MEP amplitude) may have neutralized the rTMS-induced decrease in MEP amplitude observed in

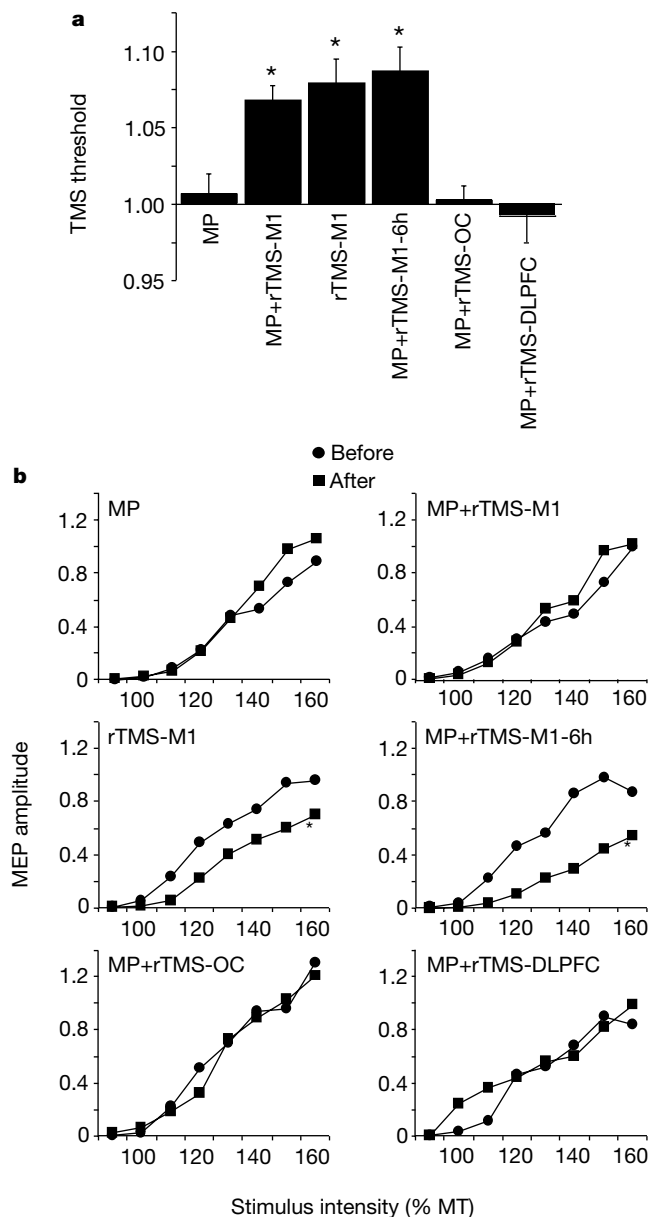


Figure 4 Interaction of focal rTMS with motor cortex excitability. Changes in TMS-evoked motor cortical output to the flexor pollicis brevis muscle. **a**, Changes in motor threshold (MT). Each bar represents the change in (normalized) mean MT for each condition, which is expressed as quotient (before/after rTMS). A significant increase in MT was seen after rTMS of M1 (MP+rTMS-M1, rTMS-M1 and MP+rTMS-M1-6h experiments). Error bars are s.e.m. Asterisk indicates $P < 0.05$. **b**, Changes in MEP amplitude. Each symbol represents the mean MEP amplitude (normalized to the maximum MEP amplitude before rTMS) for a given stimulus intensity (range 90–160% MT) before and after rTMS, and before and after the 15-min rest in the MP experiment, respectively. A significant ($P < 0.05$) depression in MEP amplitude was seen after rTMS of M1 in the rTMS-M1 and MP+rTMS-M1-6h groups but not in the MP+rTMS-M1 group.

the rTMS-M1 and rTMS-M1-6h experiments. The depressive effects of 1-Hz rTMS on M1 could reflect short-term depression or long-term depression¹², two common forms of synaptic plasticity that have been described, among other cortical areas, in M1^{12,13}.

Another important question is whether M1 was affected exclusively. One consideration is whether the TMS affected other regions by current spread. This seems unlikely because of the rapid decline in effective stimulation field strength, which is a cubic function of the distance from the centre of stimulation. Even the premotor cortex was shown not to be affected in a study of rTMS over M1¹⁴. A second consideration, although it is possible to stimulate subsets of M1 focally, is whether there are remote effects of TMS via neuronal networks, as previously demonstrated¹⁵. We cannot rule this out. However, the failure of rTMS to interfere with M1 excitability and with early motor consolidation if applied to the dorsolateral prefrontal cortex and the occipital cortex is evidence that the present effects of rTMS on M1 are restricted and not a consequence of global or nonspecific influences.

Our work fits with other experiments that show an involvement of M1 in motor learning on a cellular level. Long-term potentiation¹⁶ and long-term depression¹⁷ have been demonstrated to occur in rat motor cortex. With skill learning, horizontal cortical connections are strengthened, and the amount of long-term potentiation that can be induced in trained cortex is less than in the untrained cortex, suggesting that long-term potentiation had already produced changes toward the maximal possible strength of connectivity¹⁸. The ability to induce long-term depression in such a preparation was enhanced, however, adding further support to the concept that training had driven the synapses close to the maximum¹⁹. In the monkey, when adapting to a new dynamical environment, M1 cells could change their directional tuning to suit the new environment, and such cells maintained this new pattern even when the animal re-adapted to the original environment²⁰. Moreover, for the entire neuronal population, the shift of preferred direction matched the shift required for the task²¹. Our present results reveal an essential role of M1 in early motor consolidation of newly acquired motor skills. This further establishes the M1 as a critical node in the network for storage and processing of new motor information. □

Methods

Subjects

Forty naive right-handed normal subjects (22 men, 18 women; age mean 43 yr, range 22–57 yr) gave their written informed consent to participate in the study. The study protocol was approved by the National Institutes of Neurological Disorders and Stroke Institutional Review Board. Eight subjects participated in each experiment; the subjects from the MP+rTMS-M1 experiments also participated in the MP+rTMS-OC experiment; these two experiments were conducted at least 6 d apart.

Practice task

The practice task was a metronome-paced (0.5 Hz) ballistic pinch between the index finger and the thumb of the non-dominant left hand. The subjects were seated in front of a monitor with a virtual light displayed in its centre. For every brisk pinch with an acceleration of more than 980 cm s⁻², a green light indicated successful performance 500 ms before the next go signal. A red light indicated unsuccessful performance (peak acceleration <980 cm s⁻²). The acceleration of the pinching movement, assessed during the practice sessions, was the primary outcome measure.

Behavioural measures

The primary measure, acceleration of each thumb flexion, was recorded during practising (a total of 600 movements) with a miniature high output charge piezoelectric accelerometer with integral electronics (model 25A, Isotron PE Accelerometer, 4.575 mV g⁻¹ sensitivity, Endevco), firmly fixed to the proximal phalanx of the thumb with tape. The signal was amplified by a battery-powered low-noise signal conditioner (model 4416B Isotron Signal Conditioner, Endevco) and digitized with a PCI-MIO-16E4 board (National Instruments) at a rate of 2,000 Hz. Each trial consisted of 3,000 points (1,500 ms). The first peak acceleration of each movement was analysed automatically online for immediate feedback for task performance (green or red light), and stored for further offline analysis in a computer, using a program written in LabView Graphical Programming Language (National Instruments). Maximal peak force between the left index finger and thumb was measured with a pinch gauge (model 5083; Alimed) at the

beginning and end of each experiment according to a standardized procedure described in detail elsewhere².

Stimulation and recording technique

Surface electromyogram (EMG) was recorded (bandpass, 0.1–2.5 kHz) from the left flexor pollicis brevis (FPB) muscle using conventional EMG machines (Counterpoint Electromyograph, Dantec Electronics; Viking IV, Nicolet Biomedical). A figure-8-shaped stimulation coil connected to a rapid-rate magnetic stimulator (Cadwell Laboratories; Magstim) was positioned on the scalp over the right motor cortex at the optimal site for eliciting MEPs in the FPB. Motor threshold (MT) was defined as the minimum stimulus intensity required to produce MEPs of greater than 50 µV in at least 5 of 10 trials. Peak-to-peak MEP amplitude was measured in five trials, each at stimulus intensities ranging from 90% to 160% in steps of 10% MT. M1 excitability was assessed before and after the 15-min rTMS procedure, and before and after the 15-min rest in the MP experiment. One-hertz rTMS was performed in accordance with current safety recommendations²². rTMS over M1 was done at the optimal position for the left FPB at 115% MT in two single 15-min trains (rTMS 1 and rTMS 2) after practice 1 and 2, respectively. For the MP+rTMS-OC experiment, the intersection of the coil was placed over the occipital cortex and the current in the coil directed downward. For the MP+rTMS-DLPFC experiment, the intersection of the coil was placed 4 cm anterior to the left FPB-M1²³.

Data analysis

Behavioural and TMS data were normalized to the initial (pre-intervention) measure and given a value of 1.0. For movement acceleration, the effect of motor learning was assessed with regression analysis (correlation between number of training movements versus peak acceleration). MEP amplitudes were compared by analysis of variance (ANOVA) using a model of repeated measures with condition as between-subject factor, and time and stimulus intensity as within-subject factors. All other measures were analysed with paired *t*-test comparisons. Mean peak force was calculated from ten pinches at maximal force. Mean peak acceleration was calculated from 30 peak accelerations for each practice minute, or from 30 fast finger movements in the rTMS-M1 experiment (no practice). Values were considered statistically significant if *P* < 0.05. Pre-intervention measures of task performance, learning performance and M1 excitability showed no differences between groups (ANOVA, *P* > 0.05).

Received 4 September; accepted 4 December 2001

Published online 23 January 2002; DOI 10.1038/nature712.

- Brashers-Krug, T., Shadmehr, R. & Bizzi, E. Consolidation in human motor memory. *Nature* **382**, 252–255 (1996).
- Shadmehr, R. & Holcomb, H. H. Neural correlates of motor memory consolidation. *Science* **277**, 821–825 (1997).
- Karni, A. *et al.* Functional MRI evidence for adult motor cortex plasticity during motor skill learning. *Nature* **377**, 155–158 (1995).
- Honda, M. *et al.* Dynamic cortical involvement in implicit and explicit motor sequence learning. A PET study. *Brain* **121**, 2159–2173 (1998).
- Muellerbacher, W., Ziemann, U., Boroojerdi, B., Cohen, L. & Hallett, M. Role of human primary motor cortex in rapid motor learning. *Exp. Brain Res.* **136**, 431–438 (2001).
- Jenkins, I. H., Brooks, D. J., Nixon, P. D., Frackowiak, R. S. & Passingham, R. E. Motor sequence learning: a study with positron emission tomography. *J. Neurosci.* **14**, 3775–3790 (1994).
- Nudo, R. J., Milliken, G. W., Jenkins, W. M. & Merzenich, M. M. Use-dependent alterations of movement representations in primary motor cortex of adult squirrel monkeys. *J. Neurosci.* **16**, 785–807 (1996).
- Classen, J., Liepert, J., Wise, S. P., Hallett, M. & Cohen, L. G. Rapid plasticity of human cortical movement representation induced by practice. *J. Neurophysiol.* **79**, 1117–1123 (1998).
- Wassermann, E. M. *et al.* Use and safety of a new repetitive transcranial magnetic stimulator. *Electroencephalogr. Clin. Neurophysiol.* **101**, 412–417 (1996).
- Chen, R. *et al.* Depression of motor cortex excitability by low-frequency transcranial magnetic stimulation. *Neurology* **48**, 1398–1403 (1997).
- Kosslyn, S. *et al.* The role of area 17 in visual imagery: convergent evidence from PET and rTMS. *Science* **284**, 167–170 (1999).
- Abbott, L. F., Varela, J. A., Sen, K. & Nelson, S. B. Synaptic depression and cortical gain control. *Science* **275**, 220–224 (1997).
- Hess, G. & Donoghue, J. P. Long-term potentiation and long-term depression of horizontal connections in rat motor cortex. *Acta Neurobiol. Exp. (Warsz)* **56**, 397–405 (1996).
- Gerloff, C., Corwell, B., Chen, R., Hallett, M. & Cohen, L. G. The role of the human motor cortex in the control of complex and simple finger movement sequences. *Brain* **121**, 1695–1709 (1998).
- Paus, T. *et al.* Transcranial magnetic stimulation during positron emission tomography: a new method for studying connectivity of the human cerebral cortex. *J. Neurosci.* **17**, 3178–3184 (1997).
- Hess, G., Aizenman, C. D. & Donoghue, J. P. Conditions for the induction of long-term potentiation in layer II/III horizontal connections of the rat motor cortex. *J. Neurophysiol.* **75**, 1765–1778 (1996).
- Hess, G. & Donoghue, J. P. Long-term depression of horizontal connections in rat motor cortex. *Eur. J. Neurosci.* **8**, 658–665 (1996).
- Rioullet-Pedotti, M. S., Friedman, D., Hess, G. & Donoghue, J. P. Strengthening of horizontal cortical connections following skill learning. *Nature Neurosci.* **1**, 230–234 (1998).
- Rioullet-Pedotti, M., Friedman, D. & Donoghue, J. Learning-induced LTP in neocortex. *Science* **290**, 533–536 (2000).
- Gandolfo, F., Li, C., Benda, B. J., Schioppa, C. P. & Bizzi, E. Cortical correlates of learning in monkeys adapting to a new dynamical environment. *Proc. Natl Acad. Sci. USA* **97**, 2259–2263 (2000).
- Li, C. S., Padoa-Schioppa, C. & Bizzi, E. Neuronal correlates of motor performance and motor learning in the primary motor cortex of monkeys adapting to an external force field. *Neuron* **30**, 593–607 (2001).

22. Wassermann, E. M. Risk and safety of repetitive transcranial magnetic stimulation: Report and suggested guidelines from the international workshop on the safety of repetitive transcranial magnetic stimulation. June 5–7, 1996. *Electroencephalogr. Clin. Neurophysiol.* **108**, 1–16 (1998).
23. Pascual-Leone, A., Wassermann, E. M., Grafman, J. & Hallett, M. The role of the dorsolateral prefrontal cortex in implicit procedural learning. *Exp. Brain Res.* **107**, 479–485 (1996).

Acknowledgements

We thank our subjects for their cooperation, J.-P. Ndayisaba for technical assistance, and D. G. Schoenberg for editing. W.M. was supported by the Max Kade Foundation.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.H. (e-mail: hallettm@ninds.nih.gov).

Genome shuffling leads to rapid phenotypic improvement in bacteria

Ying-Xin Zhang*, Kim Perry*, Victor A. Vinci†, Keith Powell*, Willem P. C. Stemmer* & Stephen B. del Cardayré*

* Maxisgen, 515 Galveston Drive, Redwood City, California 94063, USA

† Eli Lilly and Company, Lilly Corporate Center, Indianapolis, Indiana 46285, USA

For millennia, selective breeding, on the basis of biparental mating, has led to the successful improvement of plants and animals to meet societal needs¹. At a molecular level, DNA shuffling mimics, yet accelerates, evolutionary processes, and allows the breeding and improvement of individual genes and subgenomic DNA fragments. We describe here whole-genome shuffling; a process that combines the advantage of multi-parental crossing allowed by DNA shuffling with the recombination of entire genomes normally associated with conventional breeding. We show that recursive genomic recombination within a population of bacteria can efficiently generate combinatorial libraries of new strains. When applied to a population of phenotypically selected bacteria, many of these new strains show marked improvements in the selected phenotype. We demonstrate the use of this approach through the rapid improvement of tylosin production from *Streptomyces fradiae*. This approach has the potential to facilitate cell and metabolic engineering and provide a non-recombinant alternative to the rapid production of improved organisms.

Evolution is a continuous process of genetic variation and phenotypic selection². Recombination within a selected population amplifies the genetic diversity of the population by creating new mutant combinations, and can thereby improve the performance of individuals within the population. We are interested in the application of recombination formats for the rapid improvement of biological systems, and here describe its application to the improvement of whole-microbial genomes. Although classical breeding addresses entire genomes, it allows for recombination between only two parents per generation. In contrast, DNA shuffling addresses DNA fragments and allows for recombination between multiple parents at each generation³. The production of the resulting multi-parent 'complex progeny' is important for the marked acceleration of directed evolution realized through DNA shuffling. A practical combination of classical breeding and DNA shuffling should thus provide a means rapidly to breed populations of organisms to produce combinatorial libraries of complex progeny.

The directed evolution of microorganisms has been accomplished traditionally through the asexual process of classical strain

improvement (CSI; sequential random mutagenesis and screening), which is the leading method for the development of commercial microorganisms^{4–6}. During each cycle of CSI a population of improved mutants is identified from which the single best performer is taken forwards. As the population of improved mutants (as opposed to the single best) from each a cycle comprises a pool of useful genetic diversity, the shuffling of this population should produce a new combination of mutants of superior performance, just as it does for subgenomic fragments^{7,8} (Fig. 1). Two primary questions were addressed by this study: can a population of bacterial cells be efficiently shuffled such that significant progeny contain genetic information from more than two parents; and does such a population contain individuals having marked improvements in a desired phenotype? We chose to shuffle *Streptomyces*, as they are the primary natural producers of commercial antibiotics, have a long history of CSI, and useful recombination formats are described^{9,10}.

Although there are various means of affecting recombination between bacterial cells, protoplast fusion^{9,10} is very efficient between *Streptomyces*, resulting in recombination efficiencies greater than 0.20 (ref. 11). The efficiency of recombination between fused protoplasts of four multiply auxotrophic strains of *Streptomyces coelicolor* have been measured previously¹². The simple two-marker progeny made up 3% of the population, while the complex three- and four-marker progeny were present at only 0.04% and 0.00005% of the population, respectively. As efficient production of complex progeny increases both the sequence and functional diversity of a population, improving the representation of complex progeny within the recombined population was necessary.

We reasoned that improvement in the distribution of complex progeny could be achieved by the recursive fusion of a mixed protoplast population (see Supplementary Information). Each pooled fusion would mimic each thermal cycle of a DNA shuffling reaction, and recursive fusions should result in the efficient shuffling of the population. To test this idea, we shuffled the same set of *S. coelicolor* strains by recursive protoplast fusion (see Methods). Protoplasts from the four strains were mixed, fused, and regenerated under non-selective conditions. Spores from the regenerated protoplasts were pooled, resulting in the first fusion library (F₁). F₁

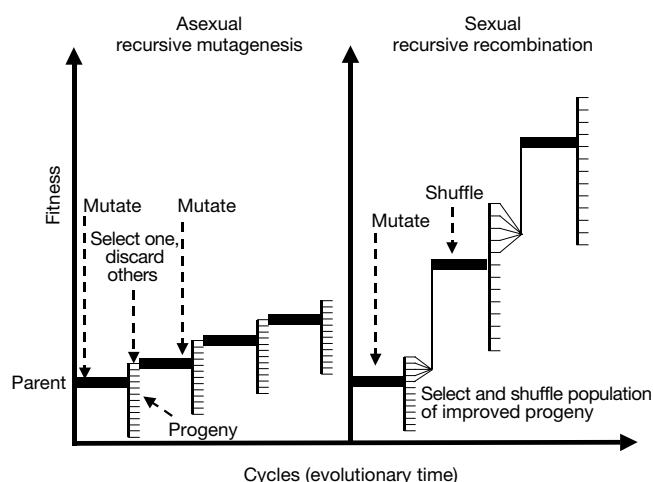


Figure 1 Asexual versus sexual evolution. Asexual evolution is the sequential process of accumulating individual mutations. Selection of the fittest results in the capture of only a single mutant. It is slow, as individuals within a population evolve alone as opposed to sharing information and evolving as a group. Genetic diversity is lost and deleterious mutations that are difficult to lose accumulate. Sex allows the information within a population to be shared. Mating within a selected population consolidates genetic information by providing a mechanism for the combination of useful mutations and the loss of deleterious mutations. Sexual evolution thus produces populations containing individuals that have a far greater fitness than their parents.

was grown as a population and used to prepare protoplasts, which were similarly fused and regenerated. This process was repeated four times resulting in the F_4 population. The distributions of two-, three- and four-marker progeny from F_1 and F_4 were determined and are shown in Table 1. Consistent with ref. 12, a single fusion resulted in the efficient production of two-marker progeny (10%) and the poor production of three-marker (0.4%) and four-marker (0.0007%) progeny. However, recursive fusion efficiently produced the two- (60%), three- (17%) and four-marker (2.5%) progeny, a $40\text{--}10^5$ -fold increase in complex progeny (the distribution of phenotypes within each population is shown in Table 2 of Supplementary Information). Thus recursive protoplast fusion was successful in the shuffling of the four *S. coelicolor* strains and could be used for the shuffling of a phenotypically selected population for the purposes of directed evolution.

Streptomyces fradiae is used for the commercial production of tylosin, a complex polyketide antibiotic¹³. Strain SF1 is a stable clone of the *S. fradiae* natural isolate, whereas SF21 produces tylosin at a high titre and is derived from SF1 through 20 cycles of CSI at Eli

Lilly (Fig. 2a). We aimed to use genome shuffling to facilitate the creation of new, high tylosin-producing strains from SF1. To generate a diverse population for genome shuffling, SF1 was subjected to one round of CSI using nitrosoguanidine (NTG) as mutagen¹⁴. A total of 22,000 individual mutants were grown in a 96-well format and screened for tylosin production (see Methods), and 11 strains were selected as a population for further evolution based on improved tylosin titres relative to SF1 (Fig. 2b). Protoplasts from each of the 11 strains were prepared, mixed in equal proportions, and recursively fused. We screened 1,000 progeny from the first cycle of shuffling for further tylosin titre improvements. Seven strains were identified that produced more tylosin than the NTG-mutated parents, and some produced tylosin at levels comparable to SF21 in our high-throughput screen (Fig. 2b). These 7 strains were subjected to an additional cycle of shuffling, and 1,000 progeny were screened. Another 7 strains with further-improved tylosin titres were identified, each producing more tylosin than SF21 under high-throughput conditions (Fig. 2b). The performance of the best two shuffled strains, GS1 and GS2, was compared to SF1 and SF21 in 250-ml shake flasks. GS1 and GS2 produce up to ninefold more tylosin than SF1 and are statistically indistinguishable in tylosin titre for SF21 (Fig. 2c).

Cellular phenotypes are by nature complex, and result from the dynamic interaction of genes and gene products distributed throughout a cell's genome. The mutations leading from SF1 to SF21, GS1 and GS2 probably lie throughout the *S. fradiae* genome and are not simply associated with the tylosin biosynthetic cluster. In an attempt to differentiate at the molecular level strains SF1, SF21, GS1 and GS2, a regulatory gene within the tylosin biosynthetic pathway believed to be involved in increased tylosin titre¹⁵ was amplified by the polymerase chain reaction and sequenced. SF21, GS1 and GS2 all had a single-nucleotide change in this gene relative to SF1; however, the mutation in GS1 and GS2 was at a different position in the gene than that in SF21. Thus, SF21, GS1 and GS2 had converged on at least one functionally similar mode of achieving improved titre, but had done so in a structurally distinct manner. At the physiological level, GS1 and GS2 are clearly distinct from SF21, as can be seen by the strain's differing morphologies. Whereas SF21 forms a small rough colony, GS1 and GS2 form round, smooth, raised colonies that resemble SF1. This difference may reflect the large mutational load SF21 carries relative to the shuffled strains.

In summary, two rounds of genome shuffling were sufficient to achieve results that had previously required 20 rounds of CSI. Theoretically, this is the difference between the asexual process of CSI and the hypersexual process of genome shuffling. Practically, this is the difference between 24,000 assays and approximately 1 year of effort, and roughly 1,000,000 assays and 20 years of effort (Vinci, V. A., personal communication). Recombination has been discussed previously for the purposes of strain improvement^{16–20}. In contrast, we demonstrate that recursive recombination within selected populations markedly increases the rate of strain evolution. By expanding the reach of shuffling technology from subgenomic fragments to whole cells and organisms, genome shuffling provides a new tool for cell and metabolic engineering that requires no sequence information or sophisticated genetic tools.

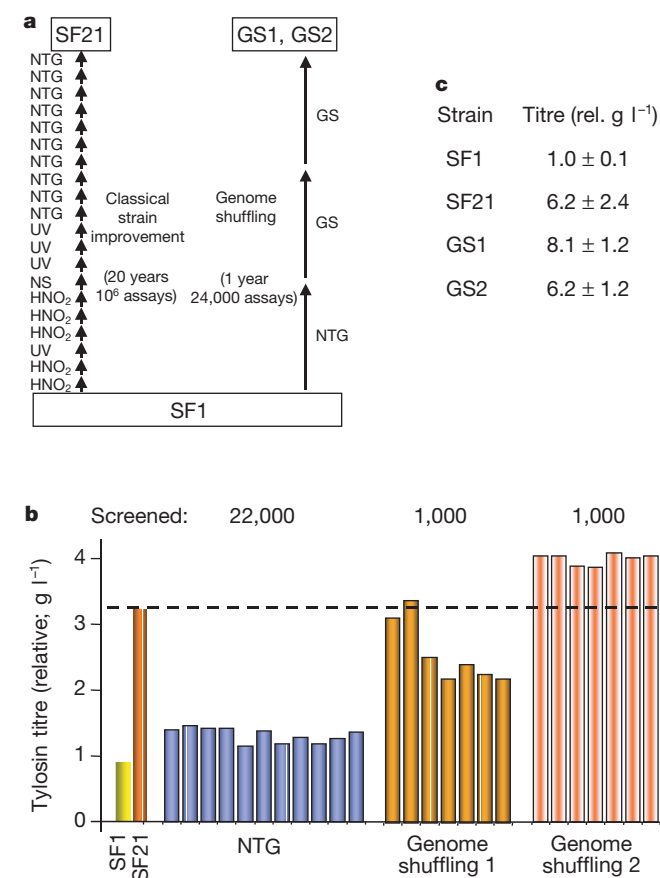


Figure 2 Genome shuffling versus classical strain improvement. **a**, Comparison of classical strain improvement to genome shuffling for production of tylosin from *S. fradiae*. SF21 was generated through 20 rounds of mutagenesis and screening from SF1. GS1 and GS2 were generated from two rounds of genome shuffling of a population of classically improved strains derived from SF1. The classical approach required twenty years and approximately a million screens, whereas genome shuffling required only a year and 24,000 screens. HNO₂, nitric acid; UV, ultraviolet irradiation; NTG, nitrosoguanidine; NS, natural selection; GS, genome shuffling. **b**, Relative production of tylosin in 300-μl, 96-well fermentations of SF1, SF21 (dotted line marks the production level of SF21); NTG, 11 NTG-improved mutants derived from 22,000 screened; genome shuffling 1, 7 strains identified from the shuffling of the NTG population, 1,000 were screened; and genome shuffling 2, the 7 strains identified from the shuffling of genome shuffling 1, 1,000 were screened. **c**, Relative production of tylosin from SF1, SF21, GS1 and GS2 from 250-ml shake flask fermentations.

Table 1 Distribution of phenotypes in a four-strain cross of *S. coelicolor*

Phenotype	Single fusion*	Recursive fusion†
Two markers	8.4%	60%
Three markers	0.73%	17%
Four markers	0.00045%	2.5%

The distribution of phenotypes from each fusion is reported in Supplementary Information.

* Phenotypes were determined from colony counts on defined medium containing 16 combinations of the four supplements (See Methods). Each phenotype is corrected for dilution and the presence of prototrophic markers, and divided by the total colonies growing on completely supplemented medium. The value shown represents the sum of the frequencies from each phenotypic class.

† The distribution of 483 individual colonies characterized for marker phenotype.

The application of natural evolutionary principles is extremely powerful in the manipulation of complex biological systems from single genes and proteins to whole organisms. The ability to rapidly evolve microbial cells provides a means to accelerate the commercialization of a variety of microbial products from pharmaceuticals to commodity chemicals, as well as create new cell lines useful for fundamental research. Although we describe recursive protoplast fusion as a useful method for the genome shuffling of *Streptomyces*, other mechanisms of genetic exchange are also useful for the genome shuffling of bacteria, and lower and higher eukaryotes²¹. As most of these formats rely on natural homologous recombination, organisms modified by these means are not considered 'genetically modified'. Genome shuffling thus represents a practical method for the rapid manipulation of the complex phenotypes of whole cells and organisms. □

Methods

Recursive protoplast fusion

Protoplasts from *S. coelicolor* M124 (*proA1 argA1 cysD18*), 2684 (*proA1 argA1 uraA1*), 2685 (*proA1 cysD18 uraA1*) and 2686 (*argA1 cysD18 uraA1*) were prepared, fused and regenerated as described previously²². Spores from the regenerated cells were pooled and used to inoculate a culture for the production of protoplasts. Protoplasts from this mixed culture were fused and regenerated. This process was repeated two additional times. In all cases rich medium supplemented with cystine, proline, arginine and uracil to minimize differential growth of individual auxotrophs was used for growth and regeneration. Fusions typically resulted in less than 1% background and greater than 10,000 regenerated colonies. The shuffling of *S. fradiae* was achieved in the same manner, except protoplast formation, fusion and regeneration were performed according to ref. 9. We carried out prototroph and tylosin analysis on random samples of pooled progeny²² (spore preparations for *S. coelicolor* or sonicated mycelia for *S. fradiae*).

HTP fermentation and tylosin assay

Clonal mycelial fragments were plated on R2YE²³ medium and grown at 30 °C for five days. Colonies were picked into 300 µl of vegetative medium (Eli Lilly) in 96-well microtitre plates. These were shaken at 30 °C and 85% relative humidity for 4 days, and 20-µl samples were transferred to 300 µl of 50% fermentation medium (Eli Lilly) in 96-deep-well plates. The plates were similarly incubated for 7 days, at which time 0.9 ml of methanol was added to each well, and the plates were centrifuged. A 20-µl sample of each supernatant was sampled and diluted into 0.1 M phosphate buffer, pH 7.0, and the tylosin content was measured spectrophotometrically as described previously²⁴. Tylosin A titre from the top 5–10% tylosin producers from each plate was more accurately determined by reverse-phase high-performance liquid chromatography²⁵. Those having improved performance were further confirmed in quadruplicate by a second fermentation starting from the original vegetative culture plate.

Shake flask analysis

Three 250-ml flasks containing 50 ml of vegetative medium were inoculated from frozen stocks. These were incubated as described above for three days, and 5-ml samples from each were used to inoculate three separate flasks of 100% fermentation medium. The 9 flasks from each strain were then incubated for 8 days and the average tylosin A content of the methanolic supernatant from each culture was determined as described above.

Received 12 September; accepted 11 December 2001.

- Burbank, L., Whitson, J., John, R., Williams, H. S. & Society, L. B. *Luther Burbank, his Methods and Discoveries and their Practical Application* (Luther Burbank, New York, 1914).
- Darwin, C. *On the Origin of Species by Natural Selection, or the Preservation of Favored Races in the Struggle for Life* (John Murray, London, 1859).
- Ness, J. E. *et al.* DNA shuffling of subgenomic sequences of subtilisin. *Nature Biotechnol.* **17**, 893–896 (1999).
- Aharonowitz, Y. & Cohen, G. The Microbiological production of pharmaceuticals. *Sci. Am.* **245**, 141–152 (1981).
- Demain, A. L. & Solomon, N. A. (eds) *Manual of Industrial Microbiology and Biotechnology* (ASM, Washington, 1986).
- Vinci, V. A. & Byng, G. in *Manual of Industrial Microbiology and Biotechnology* (eds Demain, A. L. & Davie, J. E.) 103–113 (ASM, Washington, 1999).
- Stemmer, W. P. DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution. *Proc. Natl Acad. Sci. USA* **91**, 10747–10751 (1994).
- Christians, F. C., Scapozza, L., Cramer, A., Folkers, G. & Stemmer, W. P. Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling. *Nature Biotechnol.* **17**, 259–264 (1999).
- Baltz, R. H. Genetic recombination in *Streptomyces fradiae* by protoplast fusion and cell regeneration. *J. Gen. Microbiol.* **107**, 93–102 (1978).
- Hopwood, D. A., Wright, H. M., Bibb, M. J. & Cohen, S. N. Genetic recombination through protoplast fusion in *Streptomyces*. *Nature* **268**, 171–174 (1977).
- Hopwood, D. A. & Wright, H. M. Factors affecting recombinant frequency in protoplast fusions of *Streptomyces coelicolor*. *J. Gen. Microbiol.* **111**, 137–143 (1979).

- Hopwood, D. A. & Wright, H. M. Bacterial protoplast fusion: recombination in fused protoplasts of *Streptomyces coelicolor*. *Mol. Gen. Genet.* **162**, 307–317 (1978).
- Baltz, R. H. & Seno, E. T. Genetics of *Streptomyces fradiae* and tylosin biosynthesis. *Annu. Rev. Microbiol.* **42**, 547–574 (1988).
- Baltz, R. H. in *Manual of Industrial Microbiology and Biotechnology* (Am. Soc. Microbiol., Washington, 1986).
- Stratigopoulos, G. & Cundliffe, E. Expression analysis of the tylosin-biosynthetic gene cluster: pivotal regulatory role of the *tylQ* product. *Chem. Biol.* **1**, (in the press).
- Baltz, R. H. in *Biotechnology of Antibiotics* (ed. Strohl, W. R.) 49–62 (Marcel Dekker, New York, 1997).
- Hamlyn, P. F. & Ball, C. in *Genetics of Industrial Microorganisms* 185–191 (American Society for Microbiology, Washington, 1979).
- Hopwood, D. A. in *Genetics of Industrial Microorganisms* 1–9 (American Society for Microbiology, Washington, 1979).
- Hopwood, D. A. & Chater, K. F. Fresh approaches to antibiotic production. *Phil. Trans. R. Soc. Lond. B* **290**, 313–328 (1980).
- Ikedo, H., Inoue, M. & Omura, S. Improvement of macrolide antibiotic-producing streptomycete strains by the regeneration of protoplasts. *J. Antibiot. (Tokyo)* **36**, 283–288 (1983).
- Del Cardayre, S. B. *et al.* Evolution of whole cells and organisms by recursive sequence recombination. International Patent Application No. WO 00/04190. (2000).
- Hopwood, D. A. & Wright, H. M. Bacterial protoplast fusion. *Mol. Gen. Genet.* **162**, 307–317 (1978).
- Keiser, T., Bibb, M. J., Buttner, M. J., Chater, K. F. & Hopwood, D. A. *Practical Streptomyces Genetics* (John Innes Foundation, Norwich, 2000).
- Baltz, R. H. & Seno, E. T. Properties of *Streptomyces fradiae* mutants blocked in biosynthesis of the macrolide antibiotic tylosin. *Antimicrob. Agents Chemother.* **20**, 214–225 (1981).
- Saliwanchik, R. in *Manual of Industrial Microbiology and Biotechnology* (eds Demain, A. L. & Solomon, N. A.) 410–435 (Am. Soc. Microbiol. Washington, 1986).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

The authors would like to thank D. Hopwood for the *S. coelicolor* strains, E. Lilly for *S. fradiae* SF1 and SF21, R. Baltz, E. Cundliffe, G. Huisman, V. Gavrilovic, R. Patnaik, M. Lassner, T. Cox and S. Louie for technical discussion and assistance, and the National Institute of Standards and Technology Advanced Technology Program for financial assistance.

Correspondence and requests for materials should be addressed to S.B.d.C. (e-mail: stephen.delcardayre@maxygen.com).

Vesicular restriction of synaptobrevin suggests a role for calcium in membrane fusion

Kuang Hu, Joe Carroll, Sergei Fedorovich, Colin Rickman, Andrei Sukhodub & Bazbek Davletov

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Release of neurotransmitter occurs when synaptic vesicles fuse with the plasma membrane. This neuronal exocytosis is triggered by calcium and requires three SNARE (soluble-N-ethylmaleimide-sensitive factor attachment protein receptors) proteins: synaptobrevin (also known as VAMP) on the synaptic vesicle, and syntaxin and SNAP-25 on the plasma membrane^{1–4}. Neuronal SNARE proteins form a parallel four-helix bundle that is thought to drive the fusion of opposing membranes^{5,6}. As formation of this SNARE complex in solution does not require calcium, it is not clear what function calcium has in triggering SNARE-mediated membrane fusion. We now demonstrate that whereas syntaxin and SNAP-25 in target membranes are freely available for SNARE complex formation, availability of synaptobrevin on synaptic vesicles is very limited. Calcium at micromolar concentrations triggers SNARE complex formation and fusion between synaptic vesicles and reconstituted target membranes. Although calcium does promote interaction of SNARE proteins between opposing membranes, it does not act by releasing synaptobrevin from

synaptic vesicle restriction. Rather, our data suggest a mechanism in which calcium-triggered membrane apposition enables syntaxin and SNAP-25 to engage synaptobrevin, leading to membrane fusion.

On calcium entry, readily releasable synaptic vesicles fuse with the presynaptic membrane in a rapid manner but with low probability, ensuring that neurotransmission is both fast and sustainable^{7,8}. Synaptic vesicles have been characterized in detail^{9,10} leading to identification of the neuronal calcium sensor, synaptotagmin¹¹. Membrane fusion itself requires three SNARE proteins that form a ternary complex between vesicular and target membranes⁶. It has been proposed that SNARE proteins can freely engage in a complex, which is not fusion-competent until the action of the calcium sensor^{12,13}. To better understand the behaviour of SNARE proteins and the requirement for calcium in neurotransmitter release we analysed the interaction of native synaptic vesicles with target membranes containing syntaxin and SNAP-25.

SNARE complex was purified from a detergent extract of bovine brain using anti-SNAP-25 immunochromatography (Fig. 1a, right panel). The three SNARE proteins formed SNARE complex identical to that purified from brain tissue (Fig. 1a). The rate and extent of complex formation were not enhanced in the presence of calcium (Fig. 1b).

To analyse whether SNARE complexes mediate membrane fusion we prepared fluorescently labelled syntaxin and SNAP-25 liposomes (containing 0.75% (mol/mol ratio) 7-nitro-2-1,3-benzoxadiazol-4-yl (NBD)- and rhodamine-phosphatidylethanolamine), and purified synaptic vesicles. Syntaxin and SNAP-25 were incorporated into liposomes by detergent dialysis, and high efficiency of incorporation was demonstrated by flotation of liposomes on an Optiprep gradient (Fig. 2a). Trypsin digestion in the absence and presence

of detergent demonstrated equal distribution of SNARE proteins between the outer and inner liposomal membrane leaflets (Fig. 2b). Synaptic vesicles released from synaptosomes by hypo-osmotic shock were separated from plasma membranes (containing Na⁺/K⁺-ATPase immunoreactivity) by differential centrifugation, and then floated on an Optiprep step gradient. Separation from cytosol was confirmed by western immunoblotting for vesicular synaptobrevin and cytosolic calcineurin (Fig. 2c). Uptake of glutamate by synaptic vesicles was ATP-dependent and was blocked by a proton ionophore, FCCP, showing vesicular functionality (data not shown). Electron microscopy confirmed the high purity of the synaptic vesicle preparation (Fig. 2d).

Fusion between the labelled liposomes and unlabelled synaptic vesicles is detected by unquenching of NBD fluorescence on dilution of fluorescent lipids. No fusion was observed in the absence of calcium (Fig. 2e). Addition of micromolar calcium, however, triggered fusion of synaptic vesicles with syntaxin and SNAP-25 target liposomes (Fig. 2e). Calcium also promoted SNARE complex formation (Fig. 2f), which has been shown to accompany synaptic vesicle exocytosis *in vivo*¹⁴. Botulinum toxins D and F, potent inhibitors of exocytosis, selectively remove the amino-terminal part of synaptobrevin¹⁵. Pretreatment of synaptic vesicles with either toxin abolished calcium-dependent fusion (Fig. 2g). In addition, calcium did not cause fusion of synaptic vesicles with protein-free, control liposomes (Fig. 2g). Together these data show that calcium is required to trigger SNARE-mediated fusion of synaptic vesicles with target membranes.

To understand at which stage SNARE-mediated fusion is regulated we analysed SNARE protein interactions in the absence of calcium in a single membrane model. First we tested whether syntaxin and SNAP-25, when in membranes, can engage the soluble cytoplasmic part of synaptobrevin (Fig. 3a). Approximately 50% of liposomal syntaxin and SNAP-25 formed SNARE complex with soluble synaptobrevin, consistent with full availability of syntaxin and SNAP-25 on the exterior of liposomes. On disruption of lipid bilayers by *n*-octylglucoside, intraluminal syntaxin and SNAP-25 also formed complex. Importantly, when synaptic vesicles (carrying all of synaptobrevin on their exterior) were mixed with the soluble part of syntaxin and SNAP-25, no complex could be detected unless a membrane-solubilizing detergent was added (Fig. 3b). Thus, whereas syntaxin and SNAP-25 in target membranes are freely available for SNARE complex formation, availability of synaptobrevin on synaptic vesicles is very limited. To test whether synaptobrevin, when in membranes, forms the proposed 'loose' (that is, SDS-sensitive) complex^{12,13} with syntaxin and SNAP-25, we used a botulinum toxin protection assay that uses the fact that monomeric but not complexed synaptobrevin is susceptible to proteolysis by botulinum toxin F¹⁶. Despite the presence of syntaxin and SNAP-25, vesicular synaptobrevin was fully cleaved by the botulinum toxin unless membranes were disrupted (Fig. 3c), indicating that synaptobrevin availability on native synaptic vesicles for SNARE complex formation is somehow restricted. To prove further that synaptobrevin, when in vesicular membranes, does not form a loose complex with syntaxin and SNAP-25, we analysed whether syntaxin and SNAP-25 beads can pull down (bind) vesicular synaptobrevin. No interaction of synaptobrevin with syntaxin and SNAP-25 could be detected unless membranes were disrupted by detergent (Fig. 3d). Together, these results suggest that synaptobrevin on synaptic vesicles does not form a loose complex with syntaxin and SNAP-25 in the absence of calcium.

What is the molecular basis for this synaptobrevin inhibition? Inhibition of synaptobrevin could arise from its binding to either protein or lipid on the same vesicle. Synaptophysin is thought to be the principal protein-binding partner of synaptobrevin on synaptic vesicles¹⁷. However, neurotransmission is normal in synaptophysin knockout mice¹⁸. Furthermore, synaptobrevin inhibition is disrupted by Triton X-100 (Fig. 3b), which dissolves membranes but

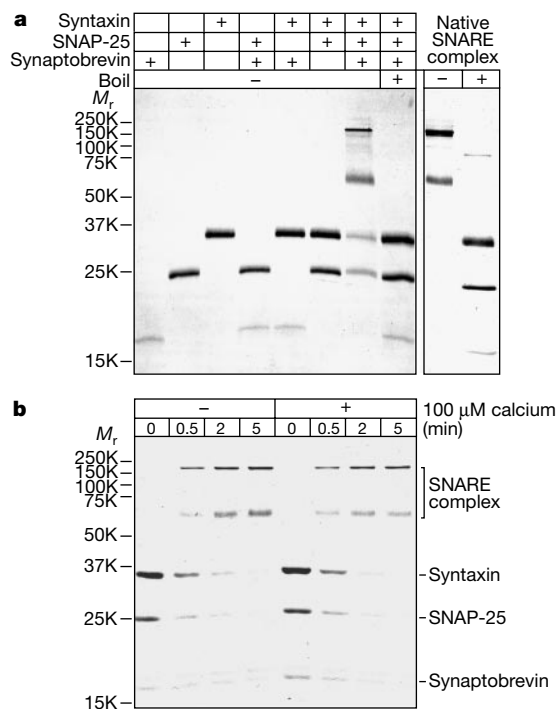


Figure 1 SNARE complex formation in solution is not enhanced by calcium. **a**, Isolated SNARE proteins form ternary complex (left panel) identical to that purified from bovine brain (right panel). SNARE proteins (30 pmol each) were incubated for 30 min and analysed by SDS-polyacrylamide gel electrophoresis (PAGE) and Coomassie staining. Samples (Boil) were boiled for 3 min to disrupt SNARE complex. **b**, SNARE proteins (10 pmol each) were incubated for the times indicated in the absence or presence of 100 μM calcium, and analysed by SDS-PAGE and Coomassie staining.

does not interfere with the reported synaptobrevin–syntaxin interaction¹⁷. Together, these findings suggest that synaptobrevin inhibition does not involve syntaxin but does require an intact vesicular lipid bilayer. Notably, synaptobrevin has been proposed to exist in an α -helical form bound to lipid membranes¹⁹. When we added soluble syntaxin and SNAP-25 to reconstituted synaptobrevin liposomes, synaptobrevin was still inhibited despite the absence of other vesicular proteins (Fig. 3e). We conclude that the vesicular lipid bilayer itself inhibits synaptobrevin.

Considering the SNARE complex does not preassemble before calcium action, how, then, does calcium activate SNARE-mediated fusion? To investigate this we incubated synaptic vesicles with soluble syntaxin and SNAP-25 or syntaxin and SNAP-25 liposomes in the absence or presence of calcium (Fig. 4a). Calcium triggers complex formation when syntaxin and SNAP-25 are in membranes, but not when they are in solution. Thus it is unlikely that calcium acts by releasing synaptobrevin from inhibition. The botulinum toxin protection assay confirms that the action of calcium in triggering SNARE complex formation strictly requires syntaxin and SNAP-25 to be in target membranes (Fig. 4b). Notably, synaptotagmin, a synaptic vesicle protein and the neuronal calcium sensor¹¹, binds trans-membranes in a calcium-dependent manner²⁰. Together with our results, this suggests that calcium may act

through synaptotagmin to bring vesicular and target membranes into close contact with each other, allowing SNAREs to form a complex and mediate fusion. To test whether calcium-dependent interaction of synaptic vesicles and target membranes takes place, we used streptavidin-coated magnetic beads to pull down protein-free biotinylated liposomes together with any bound synaptic vesicles. Synaptic vesicles bound phospholipid liposomes in response to micromolar concentrations of calcium (Fig. 4c), indicating that its addition can result in close contact between vesicular and target membranes. Finally, we incubated synaptic vesicles and syntaxin and SNAP-25 liposomes in the absence of calcium but in the presence of a low concentration of polyethyleneglycol (PEG), which promotes aggregation^{21,22}. Figure 4d shows that aggregation triggered fusion, which was blocked by botulinum toxin F. This indicates that tight apposition of membranes is sufficient to drive SNARE-mediated fusion. It is well known that hyperosmotic solutions trigger botulinum toxin-sensitive synaptic vesicle exocytosis in the absence of calcium^{23,24}. This phenomenon may be explained by the fact that hyperosmotic conditions cause membrane deformation, forcing contact between synaptic vesicles and the plasma membrane.

The presented data suggest the following mechanism by which calcium may trigger synaptic vesicle exocytosis: calcium induces

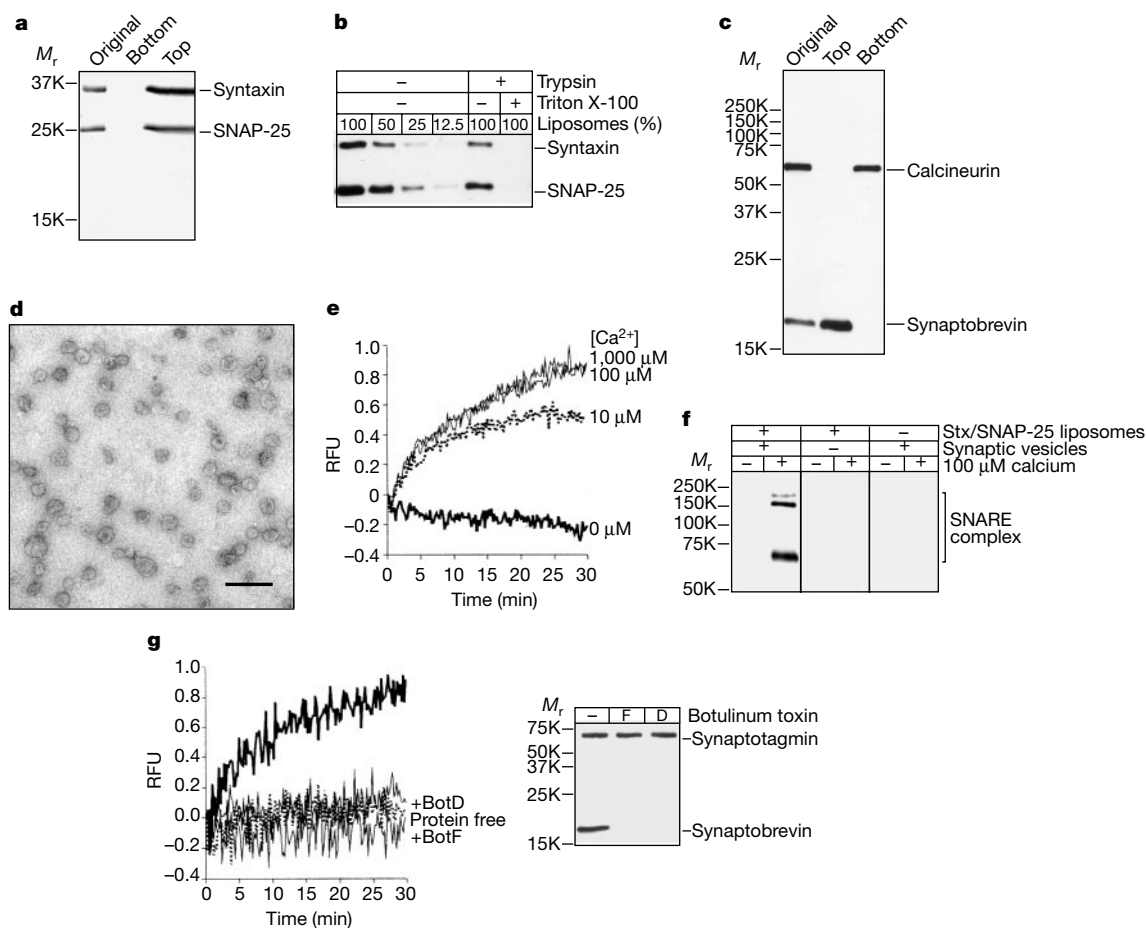


Figure 2 Synaptic vesicles fuse with target membranes in a calcium- and SNARE-dependent manner. **a**, Proteoliposomes (top) were floated from unincorporated material (bottom). Boiled samples were analysed by SDS–PAGE and Coomassie staining. **b**, SNAREs are distributed equally between outer and inner membrane leaflets. Immunoblot of proteoliposomes trypsinized in the absence or presence of Triton X-100, which dissolves membranes. **c**, Synaptic vesicles (top) floated from cytosol (bottom) were immunoblotted for cytosolic calcineurin and vesicular synaptobrevin. **d**, Electron micrograph of the synaptic vesicles. Scale bar, 200 nm. **e**, Calcium at indicated

concentrations triggers fusion of synaptic vesicles and syntaxin and SNAP-25 liposomes. All reactions contained 3.5 mM magnesium. RFU, relative fluorescence units. **f**, Synaptic vesicles and syntaxin and SNAP-25 liposomes were incubated, alone or together, for 30 min in the absence or presence of 100 μ M calcium, and immunoblotted for SNARE complex. **g**, Fusion between synaptic vesicles and syntaxin and SNAP-25 liposomes triggered by 100 μ M calcium is abolished by botulinum toxins D and F (BotD and BotF, respectively), which selectively cleave synaptobrevin; immunoblot, right panel. No fusion is observed when synaptic vesicles are mixed with protein-free liposomes.

synaptotagmin-mediated close apposition of vesicular and presynaptic membranes, allowing SNARE-driven fusion to occur rapidly and yet in a probabilistic manner (Fig. 4e). When synaptobrevin meets syntaxin and SNAP-25 heterodimer on membrane apposition, four-helix bundle formation occurs over the microsecond timescale of protein folding^{25,26}, and results in membrane fusion. We predict that the probability of fusion for individual synaptic vesicles^{7,8} can be modulated by regulating the density of preformed syntaxin and SNAP-25 heterodimer by means of Munc18/neuronal Sec1 or other factors^{2,3}. Our model is in good agreement with the properties of synaptotagmin, the neuronal

calcium sensor¹¹, which contains two calcium/phospholipid-binding domains, promotes calcium-dependent transient interaction between opposing membranes, and acts as a rapid electrostatic switch^{20,27–29}. Modulation of synaptotagmin-mediated membrane apposition may provide an additional point at which release probability can be regulated. Indeed, a single amino-acid mutation in synaptotagmin that reduced calcium/phospholipid binding resulted in decreased probability of neurotransmitter release¹¹. The present work gives mechanistic insight into the action of calcium in synaptic vesicle fusion and provides a testable hypothesis for future studies. It is plausible that in systems of cellular fusion

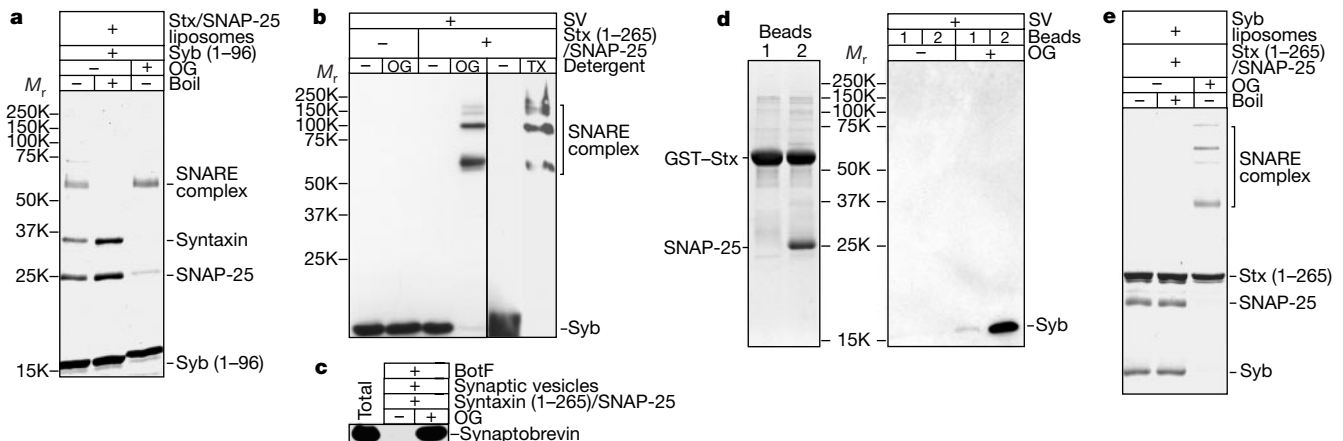


Figure 3 Availability of SNAREs in biological membranes for complex formation. All reactions were performed for 30 min in the absence of calcium. **a**, Liposomal syntaxin and SNAP-25 engage into complex with soluble synaptobrevin. OG, *n*-octylglucoside. **b**, Vesicular synaptobrevin does not complex with the cytoplasmic part of syntaxin and SNAP-25. All synaptobrevin enters complex on membrane dissolution by *n*-octylglucoside or Triton X-100 (TX). **c**, All synaptobrevin on synaptic vesicles is cleaved by botulinum

toxin F even in the presence of soluble syntaxin and SNAP-25. Membrane dissolution results in full protection of synaptobrevin (that is, SNARE complex formation). **d**, GST-syntaxin beads without (1) or with (2) pre-bound SNAP-25 (left panel; Coomassie-stained gel) were used to pull down synaptobrevin before immunoblotting of boiled samples (right panel). **e**, Synaptobrevin incorporated into liposomes does not complex with soluble syntaxin and SNAP-25.

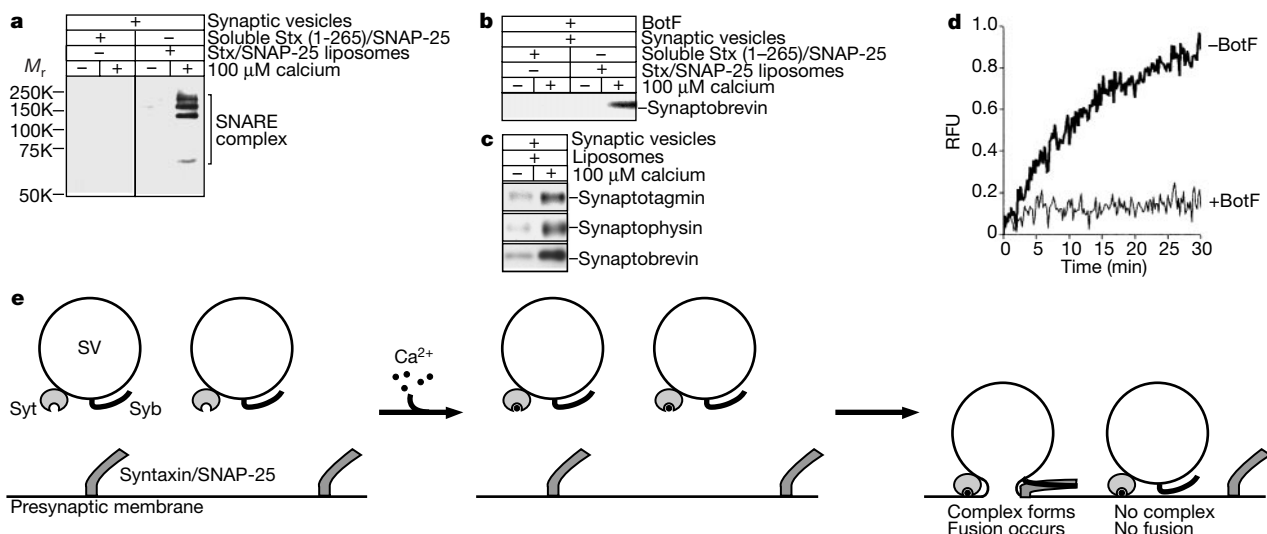


Figure 4 Apposition of vesicular and target membranes drives SNARE complex formation and fusion. **a**, **b**, Calcium action requires syntaxin and SNAP-25 to be in target membranes. Synaptic vesicles were incubated with soluble syntaxin and SNAP-25 or syntaxin and SNAP-25 liposomes in the absence or presence of 100 μM calcium (**a**). Synaptic vesicles were incubated with soluble or liposomal syntaxin and SNAP-25 before treatment with botulinum toxin F and immunoblotting of boiled samples (**b**). **c**, Calcium promotes binding of synaptic vesicles to protein-free target membranes. Liposomes pre-bound to magnetic beads were used to pull down synaptic vesicles in the absence or

presence of 100 μM calcium. Bound material was analysed for synaptic vesicle markers by immunoblotting. **d**, Aggregation of synaptic vesicles and syntaxin and SNAP-25 liposomes by 3% PEG results in fusion, which is blocked by botulinum toxin F. **e**, Model of calcium action in triggering synaptic vesicle exocytosis. Calcium triggers synaptotagmin (Syt)-mediated apposition of synaptic vesicles (SV) to the presynaptic membrane. Upon apposition, syntaxin and SNAP-25 heterodimers engage vesicular synaptobrevin (Syb) on a proportion of synaptic vesicles, consistent with the probabilistic nature of neurotransmitter release^{7,8}.

where synaptotagmin does not operate, close contact of vesicular and target membranes mediated by other factors may be instrumental in triggering SNARE-mediated membrane fusion. □

Methods

Isolation of SNARE proteins

SNARE complex was purified from bovine brain by immunoaffinity chromatography using anti-SNAP-25 SMI 81 monoclonal antibody (Sternberger Monoclonals). For preparation of the individual SNARE proteins, SNARE complex (5 mg) was heated in SDS-containing sample buffer and loaded on an 11% acrylamide gel prepared in a Mini Prep Cell (Bio-Rad). Proteins were electroeluted at 150 V and after concentration were further purified on a Superdex 200 column (Pharmacia) equilibrated in buffer A (150 mM NaCl, 20 mM HEPES, pH 7.3) containing 0.8% (w/v) *n*-octylglucoside. Plasmids encoding glutathione *S*-transferase (GST) fusion proteins of syntaxin 1A (amino acids 1–265) and synaptobrevin 2 (amino acids 1–96) were described previously¹⁶. Recombinant proteins were released from GST by incubating beads with thrombin, and passed through a Superdex 200 column equilibrated in buffer A. Detergent was removed from SNAP-25 by dialysis against buffer A. Protein concentrations were estimated using Coomassie Plus reagent (Pierce).

Preparation of proteoliposomes

Proteoliposomes were prepared essentially as described⁶. Briefly, chloroform solutions of brain phosphatidylcholine, brain phosphatidylethanolamine, brain phosphatidylserine and cholesterol (all from Avanti Polar Lipids) were mixed (48:20:12:20, mol/mol). Where necessary, 0.75% (mol/mol) NBD-dipalmitoyl phosphoethanolamine and 0.75% (mol/mol) rhodamine-dipalmitoyl phosphoethanolamine or 1% (mol/mol) Cap-biotinyl-dipalmitoyl phosphoethanolamine and [¹⁴C]dipalmitoyl phosphatidylcholine were included before drying. Dried lipids were dissolved in PBS containing 0.8% (w/v) *n*-octylglucoside. Syntaxin and SNAP-25 (1:1 mol/mol) or synaptobrevin were added where necessary, and the lipid–protein mixture (10:1 w/w) was incubated for 15 min at 37 °C before dialysis against 1 litre of PBS for 4 h at room temperature. A second round of dialysis was performed overnight at 4 °C. Dialysed sample was mixed 1:1 (v/v) with Optiprep (Invitrogen), overlaid with 30% Optiprep solution containing buffer B (140 mM potassium gluconate, 4 mM MgCl₂, 20 mM HEPES, pH 7.3), and centrifuged for 1 h at 500,000 g at 4 °C in a TLA 100.4 rotor (Beckman). Liposomes were collected from the top of the upper phase.

Isolation of synaptic vesicles

Synaptosomes were purified from rat brains as described³⁰. Synaptic vesicles were released from synaptosomes by hypo-osmotic lysis and heavy membranes were removed by centrifugation for 20 min at 34,000 g in a TLA 100.4 rotor (Beckman) at 4 °C. Supernatant was mixed 1:1 (v/v) with Optiprep, overlaid with 40% Optiprep containing buffer B, and centrifuged as for proteoliposomes. Synaptic vesicles collected from the top of the upper phase were used immediately for fusion, binding and protein assays. For electron microscopy, 3 µl of synaptic vesicles were applied to carbon-coated glow-discharged grids and stained with 2% (w/v) uranyl acetate. Electron micrographs were recorded on a Philips EM208.

Fusion assays

Fusion assays were performed in duplicate in white 96-well CliniPlates (Labsystems). In a typical experiment, 1 µl NBD/rhodamine-labelled liposomes (1 µg lipid) were mixed with 30 µl synaptic vesicles (3.5 µg protein) and 60 µl buffer B containing 2.25 mM EGTA. CaCl₂ or PEG was included as appropriate. Calcium concentrations were calculated using WEBMAXC (<http://www.stanford.edu/~cpatton/maxc.html>). Plates were then placed in a fluorimeter (Fluoroskan Ascent FL; Labsystems) equilibrated to 37 °C. NBD fluorescence was monitored with excitation and emission filters set at 460 nm and 538 nm, respectively. Initial sample fluorescence was assigned a value of zero and relative fluorescence was measured at 15-s intervals. In all figures, except Fig. 2e, spurious downward drifts were excluded by normalizing measurements to those simultaneously recorded in the absence of calcium or PEG. For botulinum toxin experiments, toxin (Calbiochem) was activated by incubation with 100 mM dithiothreitol (DTT) (20 min, 37 °C). Synaptic vesicles were incubated before use for 30 min at 37 °C with activated toxin (0.45 µg toxin per µl of synaptic vesicles) or DTT-containing buffer as a control.

Binding assay

For the synaptic vesicle binding assay, streptavidin-coated Dynabeads M-280 (Dyna) were washed in buffer B with 2% (w/v) BSA. Protein-free, radiolabelled biotinylated liposomes (see above) were mixed with the Dynabeads (0.8 mg lipid per ml Dynabeads) and agitated for 20 min at room temperature. The beads were washed twice in the same buffer containing either 1 mM EGTA or 100 µM CaCl₂, and then incubated with synaptic vesicles (0.5 µg protein) for 30 min at 37 °C. Beads were washed twice in the binding solution and bound material was solubilized in gel sample buffer. Aliquots were analysed by liquid scintillation counting for ¹⁴C content and, after lipid normalization, by western immunoblotting.

Botulinum toxin protection assay

Synaptic vesicles (0.5 µg protein) were incubated with recombinant syntaxin and SNAP-

25 (2 µg each) or syntaxin and SNAP-25 liposomes (1 µg protein) in the absence or presence of 100 µM CaCl₂ or 1% (w/w) *n*-octylglucoside for 30 min at 37 °C. Activated botulinum toxin F (2.5 µg) was then added, and incubation continued for 1 h before boiled aliquots were analysed by western immunoblotting.

Received 10 October; accepted 14 December 2001.

- Sollner, T., Bennett, M. K., Whiteheart, S. W., Scheller, R. H. & Rothman, J. E. A protein assembly-disassembly pathway *in vitro* that may correspond to sequential steps of synaptic vesicle docking, activation, and fusion. *Cell* **75**, 409–418 (1993).
- Sudhof, T. C. The synaptic vesicle cycle: a cascade of protein–protein interactions. *Nature* **375**, 645–653 (1995).
- Lin, R. C. & Scheller, R. H. Mechanisms of synaptic vesicle exocytosis. *Annu. Rev. Cell Dev. Biol.* **16**, 19–49 (2000).
- Kelly, R. B. in *Neurotransmitter Release* (ed. Bellen, H. J.) 1–33 (Oxford Univ. Press, Oxford, 1999).
- Sutton, R. B., Fasshauer, D., Jahn, R. & Brunger, A. T. Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 Å resolution. *Nature* **395**, 347–353 (1998).
- Weber, T. *et al.* SNAREpins: minimal machinery for membrane fusion. *Cell* **92**, 759–772 (1998).
- Fatt, P. & Katz, B. Spontaneous subthreshold activity at motor nerve endings. *J. Physiol. (Lond.)* **117**, 109–128 (1952).
- Schikorski, T. & Stevens, C. F. Quantitative ultrastructural analysis of hippocampal excitatory synapses. *J. Neurosci.* **17**, 5858–5867 (1997).
- Elferink, L. A. & Scheller, R. H. Synaptic vesicle proteins and regulated exocytosis. *Prog. Brain Res.* **105**, 79–85 (1995).
- Fernandez-Chacon, R. & Sudhof, T. C. Genetics of synaptic vesicle function: toward the complete functional anatomy of an organelle. *Annu. Rev. Physiol.* **61**, 753–776 (1999).
- Fernandez-Chacon, R. *et al.* Synaptotagmin I functions as a calcium regulator of release probability. *Nature* **410**, 41–49 (2001).
- Popov, S. V. & Poo, M. M. Synaptotagmin: a calcium-sensitive inhibitor of exocytosis? *Cell* **73**, 1247–1249 (1993).
- Hua, S. Y. & Charlton, M. P. Activity-dependent changes in partial VAMP complexes during neurotransmitter release. *Nature Neurosci.* **2**, 1078–1083 (1999).
- Tolar, L. A. & Pallanck, L. NSF function in neurotransmitter release involves rearrangement of the SNARE complex downstream of synaptic vesicle docking. *J. Neurosci.* **18**, 10250–10256 (1998).
- Montecucco, C. & Schiavo, G. Structure and function of tetanus and botulinum neurotoxins. *Q. Rev. Biophys.* **28**, 423–472 (1995).
- Hayashi, T. *et al.* Synaptic vesicle membrane fusion complex: action of clostridial neurotoxins on assembly. *EMBO J.* **13**, 5051–5061 (1994).
- Calakos, N. & Scheller, R. H. Vesicle-associated membrane protein and synaptophysin are associated on the synaptic vesicle. *J. Biol. Chem.* **269**, 24534–24537 (1994).
- McMahon, H. T. *et al.* Synaptophysin, a major synaptic vesicle protein, is not essential for neurotransmitter release. *Proc. Natl Acad. Sci. USA* **93**, 4760–4764 (1996).
- Quetglas, S., Leveque, C., Miquelis, R., Sato, K. & Seagar, M. Ca²⁺-dependent regulation of synaptic SNARE complex assembly via a calmodulin- and phospholipid-binding domain of synaptobrevin. *Proc. Natl Acad. Sci. USA* **97**, 9695–9700 (2000).
- Bai, J., Earles, C. A., Lewis, J. L. & Chapman, E. R. Membrane-embedded synaptotagmin penetrates *cis* or *trans* target membranes and clusters via a novel mechanism. *J. Biol. Chem.* **275**, 25427–25435 (2000).
- Viguera, A. R., Mencia, M. & Goni, F. M. Time-resolved and equilibrium measurements of the effects of poly(ethylene glycol) on small unilamellar phospholipid vesicles. *Biochemistry* **32**, 3708–3713 (1993).
- Meyuhas, D., Nir, S. & Lichtenberg, D. Aggregation of phospholipid vesicles by water-soluble polymers. *Biophys. J.* **71**, 2602–2612 (1996).
- Rosenmund, C. & Stevens, C. F. Definition of the readily releasable pool of vesicles at hippocampal synapses. *Neuron* **16**, 1197–1207 (1996).
- Capogna, M., McKinney, R. A., O'Connor, V., Gähwiler, B. H. & Thompson, S. M. Ca²⁺ or Sr²⁺ partially rescues synaptic transmission in hippocampal cultures treated with botulinum toxin A and C, but not tetanus toxin. *J. Neurosci.* **17**, 7190–7202 (1997).
- McCammon, J. A. A speed limit for protein folding. *Proc. Natl Acad. Sci. USA* **93**, 11426–11427 (1996).
- Wittung-Stafshede, P., Lee, J. C., Winkler, J. R. & Gray, H. B. Cytochrome b562 folding triggered by electron transfer: approaching the speed limit for formation of a four-helix-bundle protein. *Proc. Natl Acad. Sci. USA* **96**, 6587–6590 (1999).
- Perin, M. S. *et al.* Structural and functional conservation of synaptotagmin (p65) in *Drosophila* and humans. *J. Biol. Chem.* **266**, 615–622 (1991).
- Davletov, B., Perisic, O. & Williams, R. L. Calcium-dependent membrane penetration is a hallmark of the C2 domain of cytosolic phospholipase A2 whereas the C2A domain of synaptotagmin binds membranes electrostatically. *J. Biol. Chem.* **273**, 19093–19096 (1998).
- Zhang, X., Rizo, J. & Sudhof, T. C. Mechanism of phospholipid binding by the C2A-domain of synaptotagmin I. *Biochemistry* **37**, 12395–12403 (1998).
- Davletov, B. A. *et al.* Vesicle exocytosis stimulated by α -latrotoxin is mediated by latrophilin and requires both external and stored Ca²⁺. *EMBO J.* **17**, 3909–3920 (1998).

Acknowledgements

We thank H. McMahon for syntaxin and synaptobrevin plasmids and H. Hirling for advice on anti-SNAP-25 immunochromatography. K.H. was supported in part by the University of Cambridge MB/PhD Programme. A.S. and S.F. were supported by postdoctoral fellowships from the Royal Society and Wellcome Trust, respectively.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.D. (e-mail: baz@mrc-lmb.cam.ac.uk).

S-Cdk-dependent phosphorylation of Sld2 essential for chromosomal DNA replication in budding yeast

Hiroshi Masumoto^{*§}, Sachiko Muramatsu^{*}, Yoichiro Kamimura^{*†}
& Hiroyuki Araki^{*‡}

^{*} Division of Microbial Genetics, National Institute of Genetics, Department of Genetics, The Graduate University for Advanced Studies, Yata 1111, Mishima, Shizuoka 411-8540, Japan

[‡] PRESTO, JST, Kawaguchi, Saitama 332-0012, Japan

Cyclin-dependent protein kinases (Cdks) in eukaryotic cells work as a key enzyme at various points in the cell cycle^{1,2}. At the onset of S phase, active S-phase Cdks (S-Cdks) are essential for chromosomal DNA replication³. Although several replication proteins are phosphorylated in a Cdk-dependent manner³, the biological effects of phosphorylation of these proteins on the activation of DNA replication have not been elucidated. Here we show that Sld2 (ref. 4) (also known as Drc1; ref. 5), one of the replication proteins of budding yeast (*Saccharomyces cerevisiae*), is phosphorylated in S phase in an S-Cdk-dependent manner, and mutant Sld2 lacking all the preferred Cdk phosphorylation sites (All-A) is defective in chromosomal DNA replication. Moreover, the complex that contains, at least, Sld2 and Dpb11 (ref. 6) (the Sld2–Dpb11 complex⁴) is formed predominantly in S phase; the All-A protein is defective in this complex formation. Because this complex is suggested to be essential for chromosomal DNA replication^{4,5}, it seems likely that S-Cdk positively regulates formation of the Sld2–Dpb11 complex and, consequently, chromosomal DNA replication.

In *Saccharomyces cerevisiae*, replication origins are bound by the origin recognition complex (Orc) throughout the cell cycle^{7,8} and six minichromosome maintenance (Mcm) proteins are recruited by Cdc6 from late M to G1 phase to form the pre-replicative complex (pre-RC)⁸. Cdc45 and Sld3 join the pre-RC^{9–13}, and the subsequent activation of S-Cdks and the Cdc7–Dbf4 protein kinases facilitates the tight chromatin association of Cdc45^{11,12} and origin unwinding^{3,14,15}. Finally, DNA polymerases are recruited to origins; this step requires Dpb11 (ref. 16). In addition to positive roles in DNA replication, Cdk prevents the re-initiation of DNA replication through multiple mechanisms¹⁷ including nuclear exclusion of chromatin-unbound Mcm^{18,19}, phosphorylation and subsequent degradation of Cdc6^{20–22}, and phosphorylation of Orc¹⁷. Cdk also phosphorylates replication factor A²³ and the second-largest subunit of DNA polymerase α ²⁴, although the exact biological effects of their phosphorylation on DNA replication have not been elucidated.

Sld2, which is required for an early step of chromosomal DNA replication⁴, has 11 clustered potential Cdk phosphorylation motifs (SP and TP), 6 of which match the preferred Cdk phosphorylation motifs S/T-P-X-K/R, S/T-P-K/R and K/R-S/T-P²⁵. Because the *sld2-3* and *sld2-4* mutations (which occur in one of these motifs) are synthetically lethal with the *dpb11-1* mutation⁴, we expected that phosphorylation of Sld2 would be important for chromosomal DNA replication. We therefore examined the phosphorylation of Sld2 during the cell cycle by the mobility shift in SDS–polyacrylamide gel electrophoresis (PAGE). When cells were arrested in G1 phase by α -factor and released from G1 arrest, the Sld2 protein was observed throughout the cell cycle. Moreover, the slower-migrating forms of Sld2 appeared from early S phase (40 min) and disappeared mostly before subsequent G1 phase (120 min) (Fig. 1a). Incomplete

disappearance of the slower-migrating forms in the second G1 phase was probably due to loss of cell synchrony. Furthermore, the mobility shift of Sld2 was abolished by phosphatase treatment (Fig. 1b). These results indicate that the mobility shift of Sld2 corresponds to phosphorylation and that this phosphorylation occurs in S phase but not in G1 phase.

To investigate whether Cdk phosphorylates Sld2, we performed *in vitro* kinase assays using p13^{SUC1}-bound protein kinases²⁶. p13^{SUC1} binds specifically to Cdc28, which is the sole catalytic subunit of Cdks in budding yeast and forms a complex with various cyclins. Sld2 was phosphorylated by p13^{SUC1}-bound protein kinases (Fig. 1c). Furthermore, when we used purified human CDK kinase, the wild-type GST–Sld2 was converted to the slower-migrating form, whereas the GST–All-A protein band (see below) was not shifted (Fig. 1d). These results indicate that Cdk phosphorylates Sld2 at least *in vitro*. We then examined whether the *in vivo* phosphorylation of Sld2 in S phase depends on S-Cdk, by two methods. First, we expressed Sic1 Δ NT (non-degradable form)¹⁸, a potential inhibitor for the kinase activities of S-Cdk²⁷, in synchronized cells. In this condition, the slower-migrating forms of Sld2 did not appear (Fig. 1e, +Gal). Second, we examined Sld2 phosphorylation in cells lacking the S-cyclins Clb5 and Clb6. In these cells, Sld2 phosphorylation was inefficient and delayed after release from G1 arrest (Fig. 1f). These results strongly suggest that Sld2 phosphorylation depends on S-Cdk *in vivo*. In S phase, another protein kinase, Cdc7–Dbf4 kinase (which is essential for chromosomal DNA replication) is also active¹⁴. However, in *dbf4-1* mutant cells at conditions that are non-permissive for the activity of Cdc7–Dbf4 protein kinase, almost all the Sld2 protein was converted to the slower-migrating forms in S phase, as observed in the wild-type cells (Supplementary Information Fig. 1). This result is consistent with the idea that Sld2 is phosphorylated and converted in S phase to the slower-migrating forms by S-Cdk activity.

To examine whether Sld2 phosphorylation in S phase is essential for cell growth and chromosomal DNA replication, we constructed point mutations of the *SLD2* gene to replace a serine or threonine residue in the preferred Cdk phosphorylation motifs by an alanine residue (Fig. 2a). YYK3 cells⁴ bearing Δ *sld2* on the chromosome and the YEp195SLD2 plasmid were transformed with the plasmids harbouring the mutant alleles, and were streaked onto plates containing 5-fluoroorotic acid (FOA); this process selects cells that have lost the YEp195SLD2 plasmid. The mutant alleles bearing a single alanine substitution for any one of those motifs could substitute for the wild-type *SLD2* gene. We thus constructed the alleles bearing multiple alanine substitutions for those motifs (Fig. 2b), and introduced these to YYK3 cells. The mutant allele bearing alanine substitutions for all six serine/threonine residues (All-A) could not substitute for the wild-type *SLD2* gene (Fig. 2b). However, the mutant alleles 5A-1, 5A-5 and 5A-6, which bore the All-A substitutions except for one of three sites (S100, S208 and T241) or triple alanine substitutions for those three sites, could substitute for the wild-type *SLD2* gene (Fig. 2b). These results indicate that none of the six Cdk phosphorylation motifs is a single determinant for cell growth: they are functionally redundant.

Similar levels of Sld2 protein expressed from plasmids with wild-type, All-A and 5A-1 to -6 *SLD2* were observed in wild-type cells (data not shown), and the mutant cells harbouring the 5A-1, 5A-5 or 5A-6 allele grew as well as the wild-type cells (Supplementary Information Fig. 2). Moreover, the All-A mutant protein expressed in the wild-type cells did not show the mobility shift in S phase in SDS–PAGE (Fig. 2c) that was observed in the *in vitro* assay (Fig. 1d), and the growth defect of the All-A mutant was suppressed by the increased dosage of *DPB11* (see below). Triple aspartate substitutions (3D-3A), which may mimic the phosphorylation form, restored the cell growth of All-A (Fig. 2b; Supplementary Information Fig. 3). Although we cannot rule out the possibility that alanine substitutions affect the folding of All-A mutant protein in a

§ Present address: Imperial Cancer Research Fund, Clare Hall Laboratories, South Mimms EN6 3LD, UK

nonspecific way, these findings suggest that the All-A protein is not completely inactive but rather mimics the hypophosphorylated form. Therefore, we suggest that S-Cdk-dependent Sld2 phosphorylation is essential for cell growth.

To determine the effect of the All-A mutant allele on DNA replication, we introduced the All-A mutant plasmid to *drc1-1* cells. *drc1-1* is a temperature-sensitive allele of *SLD2* and is defective in chromosomal DNA replication at non-permissive temperatures⁵. When the *drc1-1* cells were arrested in G1 phase and released from G1 block at a non-permissive temperature, DNA content in flow cytometry did not increase for 90 min, then gradually approached 2C content (Fig. 2d), suggesting that the All-A mutant is defective in chromosomal DNA replication. This further suggests that S-Cdk-dependent Sld2 phosphorylation is required for chromosomal DNA replication.

Because Sld2 functions for chromosomal DNA replication through the Sld2-Dpb11 complex⁴, we determined whether formation of the complex requires the phosphorylation of Sld2. To detect the Sld2-Dpb11 complex, we used the strain harbouring

SLD2-10FLAG and *DPB11-9myc*. When Sld2-10Flag was immunoprecipitated with anti-Flag antibody, Dpb11-9myc co-precipitated. Conversely, when Dpb11-9myc was immunoprecipitated with anti-myc antibody, Sld2-10Flag co-precipitated. Notably, Sld2 that co-precipitated with Dpb11 showed slower migration in SDS-PAGE (Fig. 3a). We then examined complex formation during the cell cycle and found that the complex is formed predominantly in S phase (Fig. 3a). We did not examine the complex in M phase, because most Sld2-10Flag disappeared in cells arrested in M phase by nocodazole (data not shown). Moreover, when the cell extract was incubated without phosphatase inhibitors, most of the slower-migrating forms of Sld2 disappeared, and the co-immunoprecipitates of Sld2 and Dpb11 were greatly reduced (Fig. 3b). These facts suggest that the Sld2 phosphorylation in S phase is necessary for efficient formation of the complex.

We then examined whether the All-A mutant protein forms a complex with Dpb11. Because the All-A mutant is recessive (data not shown), we tried to avoid possible conflict with the wild-type Sld2 protein. For this purpose, we introduced the plasmids

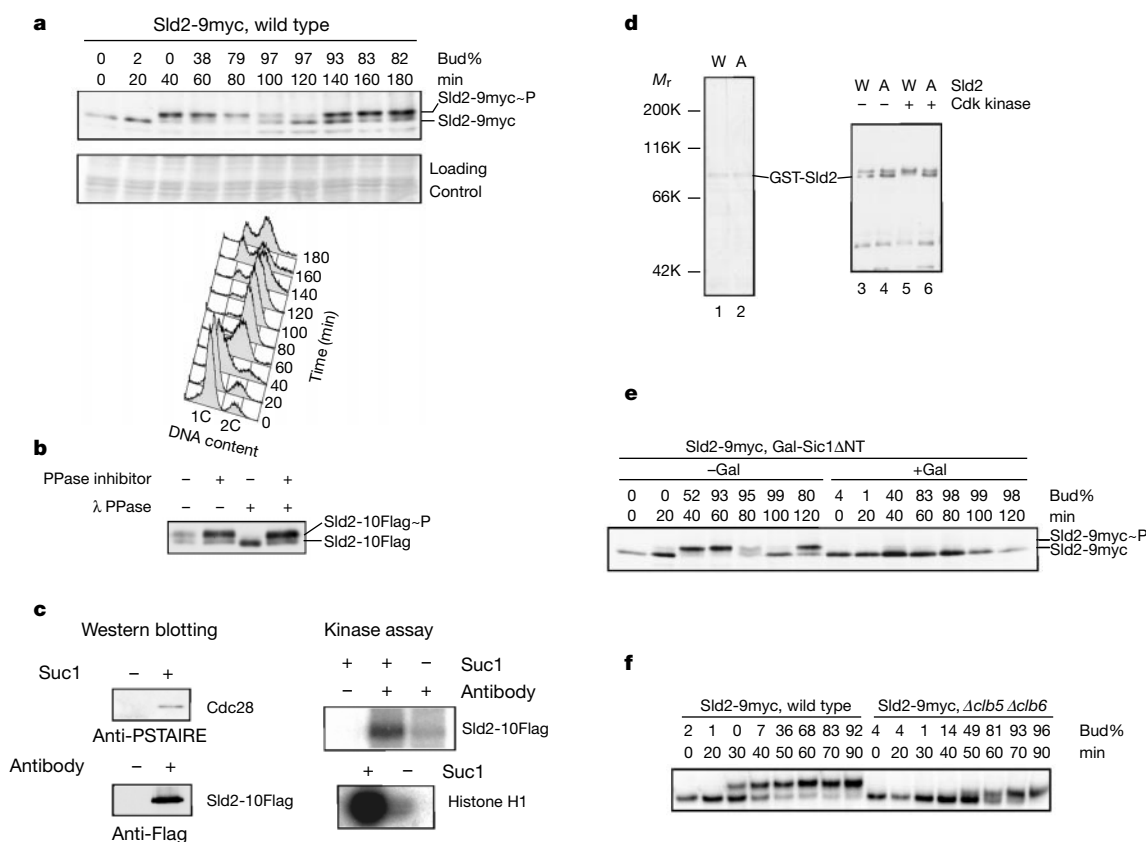


Figure 1 Sld2 is phosphorylated during the cell cycle. **a**, Cells were synchronized in G1 phase with α -factor and released in fresh YPR medium at 25 °C. Cells were withdrawn from the culture at 20-min intervals, Sld2-9myc was separated in 7.5% SDS-PAGE and detected by western blotting, and the DNA content was measured by FACS analysis. The percentage of budded cells (Bud%) is also shown. A section of the membrane between the 66K and 45K markers (relative molecular mass M_r = 66,000 and 45,000) in SDS-PAGE was stained with amido black after western blotting of the proteins as a loading control. Sld2-9myc-P indicates the slower-migrating form of Sld2-9myc, corresponding to the phosphorylation in S phase. **b**, Sld2-10Flag was immunoprecipitated from asynchronous HMS37 (*pep4::G418^r SLD2-10FLAG*) cells. The immunoprecipitates were treated (+) or not treated (-) with phosphatase (PPase) in the presence (+) or absence (-) of phosphatase inhibitor. Sld2-10Flag was separated in 7.5% SDS-PAGE and detected by immunoblotting. **c**, The proteins that bound to p13^{SUC1} beads and eluted with p13^{SUC1} protein (+) or BSA (-) were used as kinase fraction and its control and were analysed by immunoblotting with anti-PSTAIRE antibody (Santa Cruz) for Cdc28. Immunoprecipitates

with (+) or without (-) anti-Flag M2 antibody (Sigma) from cell lysates harbouring *SLD2-10FLAG* were used as substrates for kinase assay, separated in 4–20% SDS-PAGE, and radioactive proteins were detected. The shown band of radioactive protein migrated to exactly the same position as the Sld2 protein, as detected by western blotting. Kinase activity was also measured, using histone H1. **d**, The GST-Sld2 (W) and GST-All-A (A) proteins expressed in *E. coli* and purified by glutathione sepharose were run in SDS-PAGE and stained with Coomassie brilliant blue. These proteins were mixed with human CDK kinase (New England Biolabs) and detected with western blotting using anti-GST antibody (Santa Cruz). The purified GST-Sld2 protein migrates as a doublet in SDS-PAGE for unknown reasons, and the faster-migrating form shows clear mobility shift that is sensitive to Cdk inhibitors. **e**, HMS58 (*GAL1-SIC1ΔNT SLD2-9myc*) cells were released from G1 arrest either in YPR medium (-galactose) or YPRG (+galactose) at 25 °C. **f**, HMS35 (*SLD2-9myc*) and HMS45 (*clb5Δ clb6Δ SLD2-9myc*) cells were released in YPD medium at 25 °C from G1 arrest.

containing 10Flag-tagged wild-type or All-A mutant *SLD2* (Fig. 2b) to *DPB11-9myc drc1-1* mutant cells, because the Drc1-1 protein did not co-immunoprecipitate with Dpb11 in S phase at non-permissive temperatures (Fig. 3c). When the cells were arrested in S phase by hydroxyurea (HU) and the temperature then raised, Sld2-10Flag and Dpb11-9myc co-immunoprecipitated, whereas the All-A mutant protein tagged with 10-Flag and Dpb11-9myc hardly co-immunoprecipitated (Fig. 3d). We further found that the triple aspartate substitutions (Fig. 2b) in the cells arrested by expressing

Sic1ΔNT do not form a complex efficiently with Dpb11, nor do they start detectable DNA replication (data not shown). These results suggest that, although it is not all that is needed, Cdk-dependent phosphorylation is required for efficient formation of the Sld2–Dpb11 complex.

Temperature-sensitive mutants of the *SLD2* gene, *sld2-6* and *drc1-1*, are defective in formation of the Sld2–Dpb11 complex (ref. 4 and Fig. 3c) and their growth defect is suppressed by the increased dosage of *DPB11* (refs 4, 5). This suggests that low

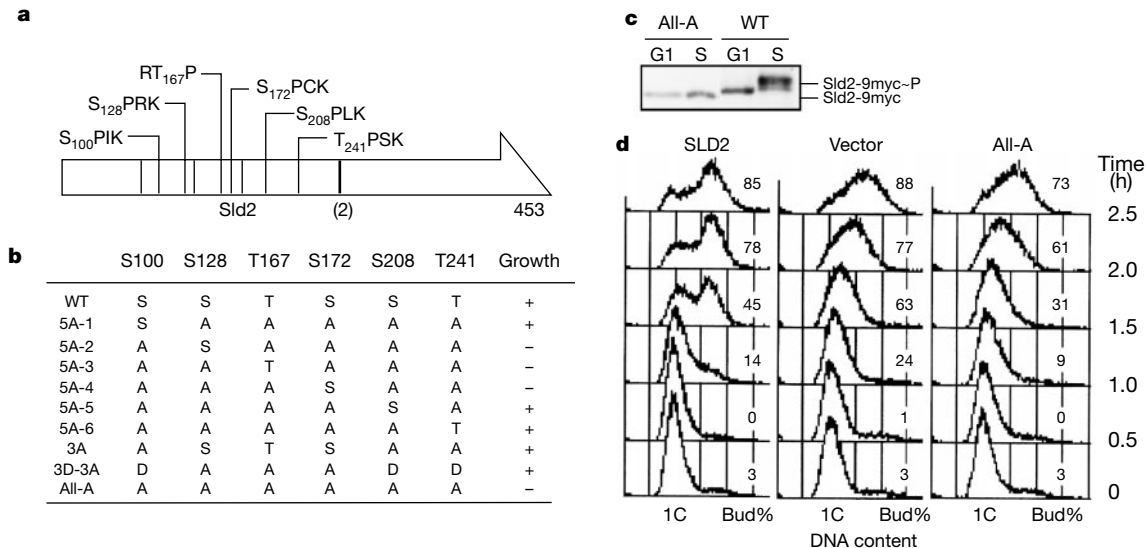


Figure 2 Phosphorylation of Sld2 in S phase is required for cell growth. **a**, Schematic presentation of CDK phosphorylation consensus sequences in Sld2. S/T-P sites are indicated by vertical lines and preferred CDK sites are depicted with their amino-acid sequences. **b**, The plasmids bearing the *sld2* mutant allele with alanine substitutions for serine/threonine of putative CDK phosphorylation motifs shown in **a** were introduced to *YK3 (sld2Δ; YEp195SLD2)* cells and the transformants were streaked onto FOA plates to examine whether they support cell growth (+) or not (–). WT, wild type. **c**, Mobility shift of

Sld2 with no substitution (WT) or six-alanine-substituted *SLD2* mutation (All-A) on plasmids was examined using a 9myc-tagged version of Sld2 in cells arrested with α -factor (G1) or 0.2 M HU (S), and then western blotted. **d**, *drc1-1* cells bearing one of YCplac22 (vector), YCp22SLD2 (SLD2) and YCp22All-A (All-A) were arrested in G1 and released at 36 °C from G1 block. Cells were withdrawn at 30-min intervals and DNA content was measured by flow cytometry. Bud%, percentage of budded cells.

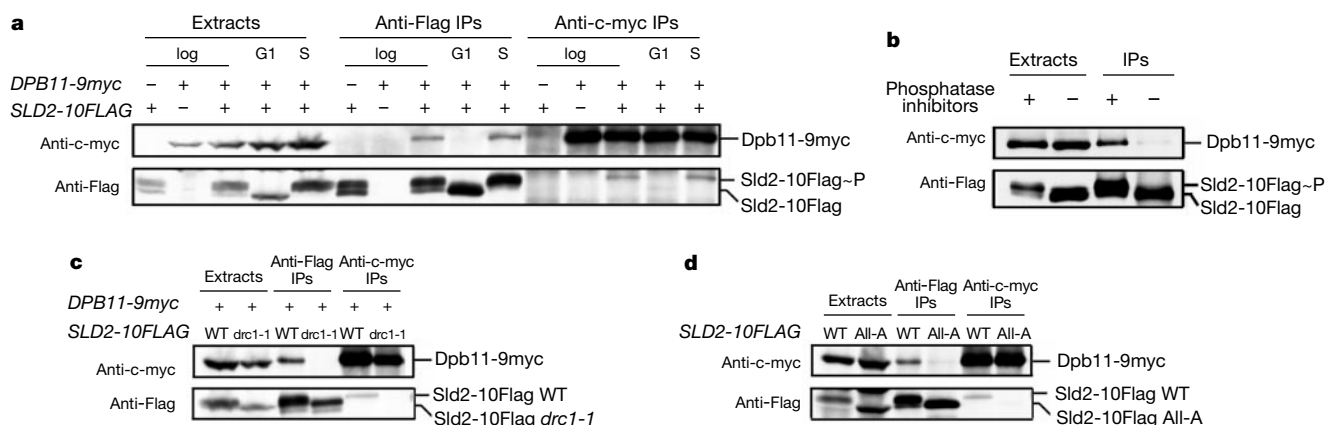


Figure 3 Phosphorylation of Sld2 is required for formation of the Sld2–Dpb11 complex. **a**, Lysates were prepared from HMS37 (*pep4::G418^r SLD2-10FLAG*), HMS40 (*pep4::G418^r DPB11-9myc*) and HMS41 (*pep4::G418^r SLD2-10FLAG DPB11-9myc*) cells in log phase (log) or arrested with α -factor (G1) or with HU (S). Sld2-10Flag and Dpb11-9myc were immunoprecipitated with anti-Flag M2 antibody (Sigma) and anti-myc 9E11 antibody (Genosys), respectively. Extracts or immunoprecipitates (IPs; corresponding to a fivefold amount of extract) were separated in 7.5% SDS–PAGE and tagged proteins were detected by immunoblotting, using anti-myc 9E10 (Roche Diagnostics) or anti-Flag M2 antibody, respectively. **b**, Lysates were prepared as described¹³ from the cells (*SLD2-10FLAG DPB11-9myc*) arrested with HU and incubated at 4 °C for 1 h with (+) or without (–) phosphatase inhibitors before immunoprecipitation

with anti-Flag M2 antibody. Immunoprecipitates correspond to a tenfold amount of extract. **c**, Co-immunoprecipitation of Sld2 and Dpb11 was performed using HMS41 and HMS70 (*pep4::G418^r drc1-1-10FLAG DPB11-9myc*) cells arrested with HU and incubated for a further 1 h at 36 °C. The mobility of Drc1-1 in SDS–PAGE is altered probably because it has four substitutions of amino acids (our unpublished result). Immunoprecipitates correspond to a fivefold amount of extract. **d**, Co-immunoprecipitation of Sld2 and Dpb11 was carried out using HMS59 (*pep4::G418^r drc1-1 DPB11-9myc*) cells harbouring YCp22SLD2-9myc (wild type, WT) or YCp22All-A-9myc (All-A); cells were arrested with HU and incubated for a further 1 h at 36 °C. Immunoprecipitates correspond to a fivefold amount of extract.

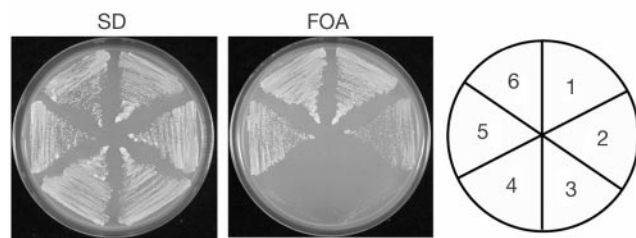


Figure 4 Increased dosage of *DPB11* suppresses the growth defect of *SLD2* All-A mutant cells. One of pRS423DPB11 (high copy) and pRS313DPB11 (low copy), and the plasmid harbouring either the wild-type or the All-A mutant *SLD2*, were introduced to *YK3* (*sld2Δ*; YEp195SLD2) cells. The transformants were grown on synthetic–glucose plates (SD) lacking uracil, histidine, leucine and tryptophan, and were streaked onto FOA plates and incubated at 25 °C for 3 d. Strains harbouring the following plasmids as well as YEp195DPB11 were streaked onto the plates: 1, YCp22SLD2 and pRS313DPB11; 2, YCp22All-A and pRS313DPB11; 3, YCp22All-A and pRS423; 5, YCp22All-A and pRS423DPB11; 6, YCp22SLD2 and pRS423DPB11.

abundance of the Sld2–Dpb11 complex in those mutant cells causes the growth defect, and that the increased dosage of *DPB11* restores the abundance of the complex and consequently restores cell growth. Because the All-A mutant is also defective in formation of the Sld2–Dpb11 complex (Fig. 3d), we examined whether the increased dosage of *DPB11* suppresses its growth defect. Both the *DPB11* plasmid and the plasmid containing the All-A mutant *SLD2* were introduced to *YK3* cells⁴ (described above) and the cells were streaked onto FOA plates. The cells lacking an extra copy of *DPB11* did not grow on FOA plates, whereas the cells with the *DPB11* plasmid did grow; furthermore, their growth level depended on the number of copies of the plasmid (Fig. 4). Because extra copies of *DPB11* do not suppress the null allele of *SLD2* (ref. 4 and Supplementary Information Fig. 4), this result indicates that the increased dosage of *DPB11* suppresses the growth defect of the All-A mutant. This result further suggests that inefficient formation of the Sld2–Dpb11 complex in the All-A mutant cells gives rise to defects in cell growth. Taken all together, it seems likely that S-Cdk-dependent phosphorylation enhances the formation of the Sld2–Dpb11 complex and thus positively regulates DNA replication.

S-Cdk-dependent phosphorylation of replication proteins has been believed to be important for activating DNA replication for a long time³ without proper demonstration. This is because real *in vivo* targets of S-Cdk to activate DNA replication have not been known. We have shown here several lines of evidence that the replication protein Sld2 is a target of S-Cdk-dependent phosphorylation, and that this phosphorylation is essential for chromosomal DNA replication. This is also true in fission yeast, because phosphorylation of SpDrc1, a possible homologue of Sld2, is essential for cell growth and DNA replication (E. Noguchi, P. Shanahan, C. Noguchi and P. Russell, personal communication). Previous studies showed that Sld2 executes its essential function with Dpb11 (refs 4, 5), which is required for the origin association of DNA polymerases at the initiation step of DNA replication¹⁶. Therefore, it is likely that S-Cdk positively regulates the origin association of DNA polymerases through the formation of the Sld2–Dpb11 complex. □

Methods

Plasmids, strains and media

pRS423DPB11 and pRS313DPB11 were constructed by inserting a *HindIII*–*Sall* *DPB11* fragment of YEp195DPB11 (ref. 6) into the *EcoRV* site of pRS423 (ref. 28) and pRS313 (ref. 29), respectively. Point mutations of *SLD2* were created on YCp22SLD2 (ref. 4) by PCR reactions using designed primers and Pfu polymerase. All the fragments obtained by PCR were sequenced to confirm no misincorporation in reactions. All yeast strains were derived from W303-1A. *SLD2*, *DPB11* and *POL2* were tagged with multiple Myc, haemagglutinin (HA) or Flag epitopes at their carboxy termini as described¹⁶. All epitope-tagged proteins are functional: the strains grew as well as the wild-type. YPD and YPR (2% raffinose instead of glucose in YPD and 40 µg ml^{−1} adenine) were used for cultivation, except YPRG

(YPR+2% galactose) was used for the induction of the *GAL1* promoter. For G1- or S-phase arrest, 30 ng ml^{−1} α-factor or 0.2 M HU was used, respectively.

Immunoprecipitation assay

Cell lysates were prepared as described³⁰. Dyna beads protein A (Dyna) pre-coated with 1 µg of anti-myc 9E11 antibody (Genosys) or 0.88 µg of anti-Flag M2 antibody (Sigma) were mixed with cell lysate (equivalent to 2 × 10⁸ cells) and rotated for 1 h. Beads were washed three times with cold B100 (50 mM HEPES-KOH buffer at pH 7.5, 5 mM magnesium acetate, 10% glycerol, 100 mM potassium acetate, 0.1% TritonX-100, 10 mM NaF, 20 mM β-glycerophosphate, 1 × Complete (Roche Diagnostics), 1 mM phenylmethylsulphonyl fluoride (PMSF) and 5 mg ml^{−1} BSA; the name B100 indicates that the buffer contained 100 mM potassium acetate); co-immunoprecipitates were then subjected to western blotting. For longer incubation in the absence of phosphatase inhibitors, cell lysates were prepared and the proteins were immunoprecipitated as previously described¹³.

Phosphatase treatment

Immunoprecipitated Sld2-10Flag was treated with λ phosphatase (40 U) (New England Biolabs) with an attached buffer system in the presence or absence of phosphatase inhibitors (60 mM β-glycerophosphate, 1 mM NaVO₃ and 50 mM NaF) at 30 °C for 20 min.

Kinase assay

The kinase assay was conducted according to a described method²⁶. Asynchronous cells were collected, resuspended in lysis buffer (50 mM Tris-HCl at pH 7.5, 150 mM NaCl, 0.1% NP-40, 5 mM EDTA, 5 mM EGTA, 0.2 mM Na₃VO₄, 50 mM KF, 20 mM β-glycerophosphate, 1 × Complete and 1 mM PMSF) and disrupted with an equal volume of acid-washed glass beads. After the cell lysate was clarified by centrifugation, 300 µg of lysate was mixed with 15 µl of p13^{sup}UCI-conjugated agarose beads (Oncogene) and rotated for 1 h. Beads were washed three times with lysis buffer and twice with kinase assay buffer (25 mM MOPS (3-*N*-morpholino-propanesulphonic acid)-KOH at pH 7.5, 15 mM MgCl₂, 0.1% NP-40, 1 mM EGTA and 0.1 mM Na₃VO₄), followed by incubation with 30 µl of 1 mg ml^{−1} p13^{sup}UCI at 4 °C for 30 min. We used the eluted fraction as a kinase fraction. From G1-arrested cells, Sld2-10Flag was immunoprecipitated with anti-Flag antibody conjugated to Dyna beads and used as a substrate for a kinase reaction. When H1 was used as a substrate of kinase assay, 1 mg ml^{−1} of histone (Sigma) was used. In the kinase reaction, immunoprecipitated Sld2-10Flag (equivalent to 2 × 10⁸ cells) were mixed with 30 µl of kinase assay buffer containing 10 µl of a kinase fraction and 30 µCi [γ-³²P]ATP. After being incubated at 37 °C for 20 min, reactions were subjected to SDS–PAGE. Incorporation of ³²P into the Sld2-10Flag protein and the histone H1 protein was detected by the BAS2000 system (Fuji film).

Received 15 November 2001; accepted 8 January 2002.

Published online 23 January 2002; DOI 10.1038/nature713.

- Nasmyth, K. At the heart of the budding yeast cell cycle. *Trends Genet.* **12**, 405–412 (1996).
- Nurse, P. A long twentieth century of the cell cycle and beyond. *Cell* **100**, 71–78 (2000).
- Kelly, T. J. & Brown, G. W. Regulation of chromosome replication. *Annu. Rev. Biochem.* **69**, 829–880 (2000).
- Kamimura, Y., Masumoto, H., Sugino, A. & Araki, H. Sld2, which interacts with Dpb11 in *Saccharomyces cerevisiae*, is required for chromosomal DNA replication. *Mol. Cell. Biol.* **18**, 6102–6109 (1998).
- Wong, H. & Elledge, S. J. *DRCl*, DNA replication and checkpoint protein 1, functions with *DPB11* to control DNA replication and the S-phase checkpoint in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **96**, 3824–3829 (1999).
- Araki, H., Leem, S.-H., Phongdara, A. & Sugino, A. Dpb11, which interacts with DNA polymerase II (ε) in *Saccharomyces cerevisiae*, has a dual role in S-phase progression and at a cell cycle checkpoint. *Proc. Natl Acad. Sci. USA* **92**, 11791–11795 (1995).
- Bell, S. P. & Stillman, B. ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature* **357**, 128–134 (1992).
- Diffley, J. F. X., Cocker, J. H., Dowell, S. J., Harwood, J. & Rowley, A. Stepwise assembly of initiation complexes at budding yeast replication origins during the cell cycle. *J. Cell Sci. (Suppl.)* **19**, 67–72 (1995).
- Aparicio, O. M., Weinstein, D. M. & Bell, S. P. Components and dynamics of DNA replication complexes in *S. cerevisiae*: Redistribution of MCM proteins and Cdc45p during S phase. *Cell* **91**, 59–69 (1997).
- Aparicio, O. M., Stout, A. M. & Bell, S. P. Differential assembly of Cdc45p and DNA polymerases at early and late origins of DNA replication. *Proc. Natl Acad. Sci. USA* **96**, 9130–9135 (1999).
- Zou, L. & Stillman, B. Formation of a preinitiation complex by S-phase cyclin CDK-dependent loading of Cdc45p onto chromatin. *Science* **280**, 593–596 (1998).
- Zou, L. & Stillman, B. Assembly of a complex containing Cdc45p, replication protein A, and Mcm2p at replication origins controlled by S-phase cyclin dependent kinases and Cdc7p-Dbf4p kinase. *Mol. Cell. Biol.* **20**, 3086–3096 (2000).
- Kamimura, Y., Tak, Y.-S., Sugino, A. & Araki, H. Sld3, which interacts with Cdc45 (Sld4), functions for chromosomal DNA replication in *Saccharomyces cerevisiae*. *EMBO J.* **20**, 2097–2107 (2001).
- Johnston, L. H., Masai, H. & Sugino, A. First the CDKs, now the DDKs. *Trends Cell Biol.* **9**, 249–252 (1999).
- Tanaka, T. & Nasmyth, K. Association of RPA, with chromosomal replication origins requires an Mcm protein, and is regulated by Rad53, and cyclin- and Dbf4-dependent kinases. *EMBO J.* **17**, 5182–5191 (1998).
- Masumoto, H., Sugino, A. & Araki, H. Dpb11 controls the association between DNA polymerases α and ε and the autonomously replicating sequence region of budding yeast. *Mol. Cell. Biol.* **20**, 2809–2817 (2000).

17. Nguyen, V. Q., Co, C. & Li, J. J. Cyclin-dependent kinases prevent DNA re-replication through multiple mechanisms. *Nature* **411**, 1068–1073 (2001).
18. Labib, K., Diffley, J. F. X. & Kearsey, S. E. G1-phase and B-type cyclins exclude the DNA-replication factor Mcm4 from the nucleus. *Nature Cell Biol.* **1**, 415–422 (1999).
19. Nguyen, V. Q., Co, C., Irie, K. & Li, J. J. Clb/Cdc28 kinases promote nuclear export of the replication initiator proteins Mcm2–7. *Curr. Biol.* **10**, 195–205 (2000).
20. Elasser, S., Chi, Y., Yang, P. & Campbell, J. L. Phosphorylation controls timing of Cdc6p destruction: A biochemical analysis. *Mol. Biol. Cell* **10**, 3263–3277 (1999).
21. Drury, L. S., Perkins, G. & Diffley, J. F. X. The cyclin-dependent kinase Cdc28p regulates distinct modes of Cdc6p proteolysis during the budding yeast cell cycle. *Curr. Biol.* **10**, 231–240 (2000).
22. Calzada, A., Sanchez, M., Sanchez, E. & Bueno, A. The stability of the Cdc6 protein is regulated by cyclin-dependent kinase/cyclin B complexes in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **275**, 9734–9741 (2000).
23. Din, S., Brill, S. J., Fairman, M. P. & Stillman, B. Cell-cycle-regulated phosphorylation of DNA replication factor A from human and yeast cells. *Genes Dev.* **4**, 968–977 (1990).
24. Foiani, M., Liberi, G., Lucchini, G. & Plevani, P. Cell cycle-dependent phosphorylation and dephosphorylation of the yeast DNA polymerase α -primase B subunit. *Mol. Cell. Biol.* **15**, 883–891 (1995).
25. Pearson, R. B. & Kemp, B. E. Protein kinase phosphorylation site sequences and consensus specificity motifs: Tabulations. *Methods Enzymol.* **200**, 62–81 (1991).
26. Pondaven, P., Meijer, L. & Beach, D. Activation of M-phase-specific histone H1 kinase by modification of the phosphorylation of its p34cdc2 and cyclin components. *Genes Dev.* **4**, 9–17 (1990).
27. Schwob, E., Böhm, T., Mendenhall, M. D. & Nasmyth, K. The B-type cyclin kinase inhibitor p40^{Stc1} controls the G1 to S transition in *S. cerevisiae*. *Cell* **79**, 233–244 (1994).
28. Christianson, T. W., Sikorski, R. S., Dante, M., Shero, J. H. & Hieter, P. Multifunctional yeast high-copy-number shuttle vectors. *Gene* **110**, 119–122 (1992).
29. Sikorski, R. S. & Hieter, P. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics* **122**, 19–27 (1989).
30. Zachariae, W. *et al.* Mass spectrometric analysis of the anaphase-promoting complex from yeast: Identification of a subunit related to cullins. *Science* **279**, 1216–1219 (1998).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank T. Iida, J. F. X. Diffley, K. Labib, H. Wong, S. Elledge, K. Sugimoto and H. Masukata for strains and plasmids; T. Kishi and H. Seino for PSTAIRE antibody and discussions; P. Russell for information regarding the unpublished results; and J. F. X. Diffley, K. Labib, H. Masukata, H. Takisawa and A. Sugino for critical reading of the manuscript. This study is partially supported by grants-in-aid from the Ministry of Education, Culture, Sports, Science and Technology, Japan, to H.A. H.M. was a postdoctoral fellow in the National Institute of Genetics.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.A. (e-mail: hiaraki@lab.nig.ac.jp).

An exonucleolytic activity of human apurinic/apyrimidinic endonuclease on 3' mispaired DNA

Kai-Ming Chou & Yung-Chi Cheng

Department of Pharmacology, Yale University School of Medicine, 333 Cedar Street, Sterling Hall of Medicine B-315, New Haven, Connecticut 06520, USA

Human apurinic/apyrimidinic endonuclease (APE1) is an essential enzyme in DNA base excision repair that cuts the DNA backbone immediately adjacent to the 5' side of abasic sites to facilitate repair synthesis by DNA polymerase β (ref. 1). Mice lacking the murine homologue of APE1 die at an early embryonic stage². Here we report that APE1 has a DNA exonuclease activity on mismatched deoxyribonucleotides at the 3' termini of nicked or gapped DNA molecules. The efficiency of this activity is inversely proportional to the gap size in DNA. In a base excision repair system reconstituted *in vitro*, the rejoining of nicked

mismatched DNA depended on the presence of APE1, indicating that APE1 may increase the fidelity of base excision repair and may represent a new 3' mispaired DNA repair mechanism. The exonuclease activity of APE1 can remove the anti-HIV nucleoside analogues 3'-azido-3'-deoxythymidine and 2',3'-dideoxy-2',3'-dideoxythymidine from DNA, suggesting that APE1 might have an impact on the therapeutic index of antiviral compounds in this category.

The DNA base excision repair (BER)³ system contributes to maintenance of the genetic integrity of the cell. BER is initiated by the action of DNA glycosylases on damaged bases generating apurinic/apyrimidinic (AP) sites³. These non-instructional lesions, which block replication and cause cell death⁴, can also be generated by non-enzymatic hydrolytic depurination. APE1 is the key enzyme that repairs AP sites through endonuclease action immediately adjacent to the 5' side of AP sites, generating 5' deoxyribose-5-phosphate and nucleotide 3' hydroxyl-terminal residues¹.

Two pathways have been described for completing the BER process: long-patch and short-patch repair³. In short-patch repair, which is considered to be the dominant pathway for BER, DNA polymerase β inserts a nucleotide to fill in the gap and releases 5' deoxyribose-5-phosphate through its intrinsic lyase activity⁵; DNA ligase I (ref. 3) or DNA ligase III/XRCC1 (ref. 3) then seals the nick. However, DNA polymerase β lacks a proofreading 3' $\rightarrow$ 5' exonuclease activity⁶ and is prone to error, misincorporating 1 nucleotide in 4,000 (ref. 3). The misincorporation of nucleotides during DNA replication or repair processes results in mutagenesis and carcinogenesis⁷. The relatively low fidelity of DNA polymerase β , combined with a conservative estimate of 20,000 AP sites produced per cell per day³, can result potentially in several mutations per day per genome.

In addition to its endonuclease activity, APE1 has 3' $\rightarrow$ 5' DNA exonuclease, 3' phosphatase, 3' phosphodiesterase and RNaseH activities. The 3' $\rightarrow$ 5' exonuclease was considered to be insignificant biologically, because it is 2–4 orders of magnitude less active than the AP endonuclease^{8,9}. However, we previously identified¹⁰ APE1 as the principal exonuclease that removes the stereochemically unnatural L-configuration anticancer nucleoside analogue, β -L-dioxolane-cytidine (L-OddC, BCH-4556; Troxacitabine)¹¹, and other L-configuration nucleoside analogues from the 3' terminus of DNA in leukaemic cell nuclei.

To examine the effect of base–base hydrogen bonding on the exonuclease activity of APE1, we carried out a time course experiment using various combinations of recessed DNA with a 3' mispair as substrates (Fig. 1a; and Supplementary Information Fig. 1b, d). The initial rates of the 3' mispair removal are summarized in Table 1. APE1 removed 3' mismatched nucleotides at least 50 times more efficiently than those matched correctly. The same result was observed with APE1 purified from the nuclei of acute lymphocytic leukaemia cells¹⁰.

As APE1 has been proposed to have a key role in coordinating the serial steps of BER^{12,13}, we studied whether the APE1 exonuclease activity might have a role in correcting mispairs introduced by the error-prone DNA polymerase β (ref. 3) during BER by using double-stranded DNA oligonucleotides containing a mispair at the 3' terminus of a single-strand break as substrates. The mismatched nucleotide was removed more efficiently from nicked DNA than from recessed DNA (Fig. 1b; and Supplementary Information Fig. 1c, e). More notably, the reaction initial rates of APE1 exonuclease showed a significant preference (160-fold) for mispaired substrates than for matched nucleotides (Table 1).

The 3' mispair exonuclease activity and the endonuclease activity on furan DNA (an abasic DNA analogue) substrate chromatographically co-eluted during APE1 purification, which, together with the finding that the double site-directed APE1 mutant lacks both AP endonuclease and exonuclease activity, confirmed that these two activities are properties of APE1 (Supplementary Information Figs 2 and 3).

Previous studies have indicated that APE1 requires a minimum of four base pairs 5' to the AP site, and three base pairs 3' to the site, for endonuclease activity¹. Thus, the difference in mismatched nucleotide removal efficiencies between the nicked and recessed DNA might be due to differences in the stability of APE1–DNA binding. We therefore tested mispairs at the 3' end of a nick or gaps of variable sizes as substrates. APE1 showed similar efficiency with nicked DNA and with DNA containing a single nucleotide gap, and showed the lowest efficiency for recessed DNA (Fig. 2; and Supplementary Information Fig. 4). The efficiency of 3' mispair removal decreased as the gap size increased. By contrast, DNA substrates with different combinations of internal mismatches (that is, in the absence of a nick or a gap) were not substrates of APE1 (data not shown).

We used an *in vitro* reconstituted system to investigate further the potential role of APE1 in correcting DNA polymerase β errors during BER. It has been reported that DNA ligase I is inefficient in rejoining nicked DNA when the nucleotide pair at the 3' terminus of the nick is mismatched¹⁴. In the absence of APE1, we observed less than 10% of ligated product with mispaired DNA, compared with at least 98% of product with the correct pair at the 3' terminus (Fig. 3a). The addition of APE1 (up to 270 pM) had no impact on the ligation of correctly paired DNA substrates, but increased the

Table 1 Specific exonuclease activities of APE1 on different DNA substrates

Base pairs	Nicked DNA	Recessed DNA
A/G	10.0	4.8
T/G	32.0	8.0
G/G	10.4	8.0
C/G	0.2	0.1

The assay conditions are the same as Fig. 1. Reaction initial rates (DNA removed in nM min⁻¹) from the time course experiments (Fig. 1) were used to determine the specific activity of APE1 exonuclease on different DNA substrates (DNA removed in pM per min per pM protein).

efficiency of ligating the 3' mispaired substrates in a concentration-dependent manner (Fig. 3a). These data support a role for APE1 in removing 3' mispaired nucleotides at the nick site during short-patch BER.

We used the antiviral nucleoside analogues 3'-azido-3'-deoxythymidine (AZT) and 2',3'-dideohydro-2', 3'-dideoxythymidine (D4T) as substrates to investigate the influence of sugar geometry on the APE1 exonuclease. Because all three analogues lack a 3' hydroxyl group, no ligation product formed in the presence of DNA ligase I alone, but the addition of APE1 removed all three analogues from the 3' terminus of a nick (Fig. 3b). In the presence of APE1, ligated products from polymerase β and ligase I were observed in proportion to the amount of APE1 used (Fig. 3b). Repair efficiency varied slightly among the three nucleoside analogues, perhaps owing to the different sugar geometry.

Previous studies have indicated that both D4T and AZT are incorporated into DNA by cellular DNA polymerases¹⁵, and over-expression of DNA polymerase β enhances the cytotoxicity of AZT¹⁶. In contrast to cytosolic exonucleases, which remove AZT and D4T poorly from DNA¹⁷, APE1 acted efficiently against both nucleoside analogues when present in nicked DNA or gapped DNA. Given that APE1 is expressed constitutively in the cell and is distributed primarily in the nucleus, APE1 might be involved in reducing the cytotoxicity¹⁸ and improving the therapeutic index of anti-HIV compounds that belong to the chain terminator category such as AZT and D4T.

Using an APE1 "trap"¹⁹, we examined the relative contribution of APE1 to the total exonucleolytic activity on 3' mispairs of nicked DNA in crude nuclear extracts of human cells (Fig. 4a). The trap eliminated both the 3' mispair exonuclease and AP endonuclease

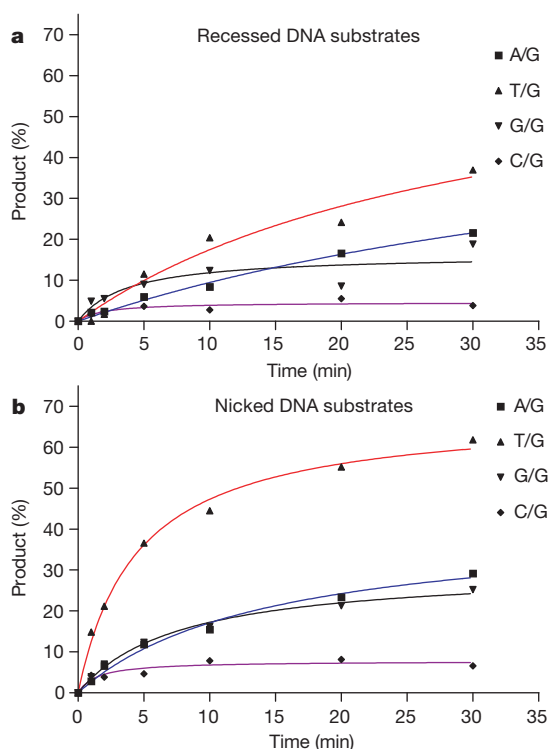


Figure 1 Rate of APE1 exonuclease activity on different DNA substrates. In each reaction, 1 nM DNA substrate was incubated with purified recombinant APE1 for various amounts of time under the standard assay conditions. For the A/G, T/G and G/G DNA substrates, 5 pM APE1 was used. For the C/G DNA substrate, 125 pM APE1 was used. The enzyme specificity of APE1 on F15 DNA was 313. **a**, APE1 exonuclease activity on 3' mispaired recessed DNA. **b**, APE1 exonuclease activity on 3' mispaired nicked DNA.

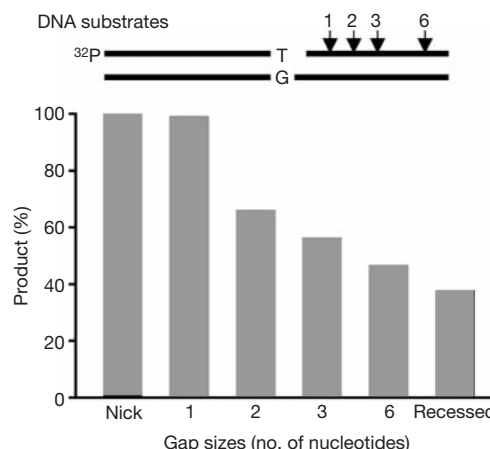
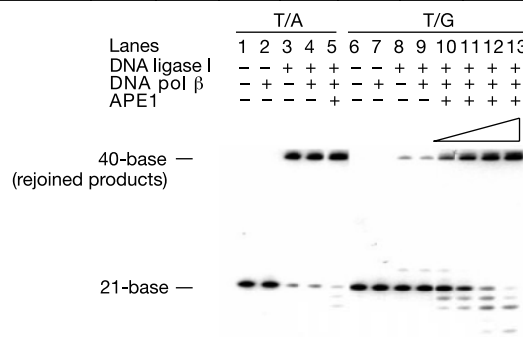


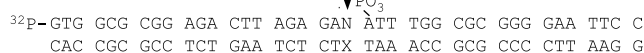
Figure 2 Activity of APE1 mismatch repair on DNA with different gap sizes. Arrows indicate the position and sizes of the gaps. For each reaction, APE1 (13.5 pM) was added to different substrates (1 nM) under the standard exonuclease assay condition at 37 °C for 30 min. The efficiencies of 3' mispair removal of different DNA substrates were adjusted using the efficiency of nicked DNA as a 100% standard. The specific activity of APE1 on F15 DNA was 257.

5'-³²P-GTG GCG CGG AGA CTT AGA GAT ^{OHPO₃} ATT TGG CGC GGG GAA TTC C
CAC CGC GCC TCT GAA TCT CTX TAA ACC GCG CCC CTT AAG G-5'

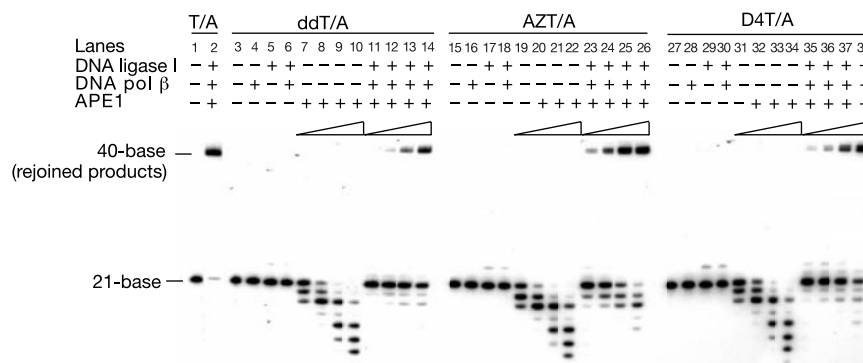
	X = A or G								
Lanes	3	4	5	8	9	10	11	12	13
Rejoining (%)	≥ 98	≥ 98	≥ 98	< 10	< 10	23 ± 2	50 ± 8	78 ± 7	95 ± 5



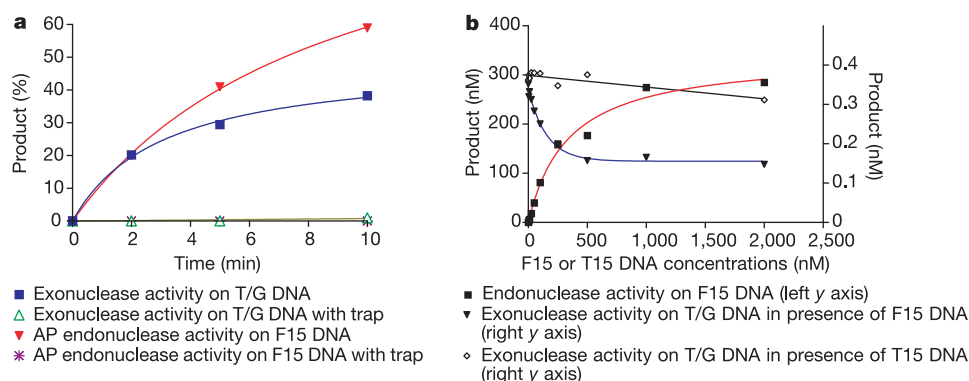
↓
PO⁻



Lanes	2	11	12	13	14	23	24	25	26	35	36	37	38
Rejoining (%)	≥ 98	< 10	< 10	17±1	44±3	< 10	23±2	56±4	65±4	< 10	< 10	28±3	64±



7–10, 11–14, 19–22, 23–26, 31–34 and 35–38 of **b**, increasing concentrations of APE1 were used (0, 2.7, 13.5, 67.5 and 270 pM, respectively). The enzyme specific activity of APE1 on the F15 DNA was 257.



nuclear extract at 37 °C under the standard exonuclease assay conditions. **b**, Inhibition of 3' mispair exonuclease activity in nuclear extract by F15 DNA. For each reaction, 62.5 ng nuclear extract, 1 nM ³²P-labelled T/G mispaired nicked DNA, 1 nM ³²P-labelled F15 DNA or undamaged T15 DNA, and increasing concentrations of unlabelled F15 or T15 DNA were incubated for 5 min at 37 °C under the standard exonuclease assay conditions.

activities in the nuclear extract. We then carried out a competition assay in which increasing concentrations of an oligonucleotide containing the AP site analogue tetrahydrofuran (F15) were added to a fixed concentration of a crude nuclear extract (Fig. 4b). Both the endonucleolytic action on F15 DNA and exonucleolytic action on a 3'-terminal mispair were monitored simultaneously. We chose F15 DNA as a competitor because it is a specific substrate for AP endonucleases, unlike aldehydic AP sites (which are also cleaved by AP lyases such as Nth1 and OGG1; ref. 20), and it is very similar or identical to natural aldehydic AP sites. The observed rate of the 3' mispair removal was inversely proportional to the concentration of F15 DNA and the rate of the AP endonuclease activity (Fig. 4b). The F15 DNA could inhibit most 3' mispair removal (~50% at 500 nM), whereas an undamaged oligonucleotide (T15) with identical sequence to the F15 DNA substrate competed very poorly (less than 3% at 1,000 nM). Similar results were observed when the assay was performed using purified APE1 instead of a nuclear extract (data not shown). Given that no other human enzyme has been shown to possess both AP-specific and exonuclease activities, and that APE1 is responsible for more than 95% of the AP endonuclease activity in the cell²¹, these data indicate that APE1 may be responsible for most of the 3' mispair exonuclease activity in the nucleus.

We found that in the reconstituted BER system, the rejoining reaction is dependent on the presence of APE1. In addition, the lower efficiency of DNA ligase I in rejoining mispaired DNA creates a window of time for APE1 to remove the mismatched nucleotide; a similar phenomenon has also been observed with DNA ligase III (ref. 22). Although other 3' → 5' DNA exonucleases such as TREX 1 have been proposed to provide a proofreading function for DNA polymerase β (refs 23, 24), APE1 is conceptually more suitable for this task as its physical interaction with DNA polymerase β has already been shown¹³. The intracellular localization of TREX 1 has not been determined clearly, and a study has indicated that TREX 1 has only a minor preference for DNA with a 3' mispair compared with its preference for matched DNA²⁵. In contrast, APE1 localizes in the nucleus (data not shown), where BER takes place.

These results have expanded the functional scope of APE1, and this versatile and important enzyme may prove to be a major contributor to the observed high fidelity of DNA repair and replication. Our study provides a new perspective from which to view the role of APE1 during early embryonic development. Moreover, the 3' → 5' DNA exonuclease activity of APE1 provides an appropriate pharmacological target for the development of antiviral and anticancer nucleoside analogues. □

Methods

Protein purification

For expression and purification of the wild-type and mutant APE1 proteins, see Supplementary Information.

Oligonucleotide substrates

Sequences of DNA substrates used were as follows. Tetrahydrofuran DNA (F15):



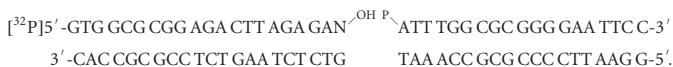
T15 DNA:



Recessed mismatched DNA (N/G, N = A, T or G):



Nicked mismatched and matched DNA (N/G, N = A, T, G or C):



Oligonucleotides with ddT-MP, AZT-MP and d4T-MP were synthesized, and their purity was confirmed as described¹⁰. The purified single-stranded oligonucleotides containing nucleoside analogues at the 3' termini (21-base) and the downstream single-stranded

oligonucleotide (19-base) were then annealed to a 40-base template to form nicked double-stranded DNA substrates.

Exonuclease/endonuclease assays

The standard exonuclease/endonuclease reaction contains 1 nM ³²P 5'-end-labelled DNA substrate and 20 mM Tris-HCl (pH 7.4) 2 mM MgCl₂, 0.5 mM EDTA, 30 mM KCl and 0.1 mg ml⁻¹ bovine serum albumin. The reactions were stopped by adding loading solution (90% formamide, 1 mM EDTA, 0.1% xylene cyanole and 0.1% bromophenol blue), and were heated at 80 °C for 3 min. We analysed samples by a 12.5% polyacrylamide gel containing 8 M urea. The gel was then dried under vacuum and subjected to autoradiography and phosphorimaging for quantification (Biorad). We calculated the percentage of removal efficiency of APE1 exonuclease by dividing the product radioactivity by the total radioactivity of substrate and product. One unit of enzyme activity is defined as 1 nM of F15 DNA cleaved in 30 min at 37 °C under the assay conditions.

In vitro reconstitution of the BER system

The reconstitution reaction was carried out under the standard exonuclease assay condition with the supplement of 30 μM dNTP, 2 mM ATP, DNA polymerase β (0.5 nM), DNA ligase I (20 nM) and increasing concentrations of APE1. We incubated the reaction mixtures for 30 min at 37 °C. The efficiency of ligation was calculated by dividing the radioactivity of rejoined product by the total radioactivity of substrate and product.

Cell culture

We grew HepG2 cells in RPMI-1640 medium (Sigma) supplied with 10% fetal bovine serum (FBS) at 95% humidity with 5% CO₂ in air at 37 °C.

Nuclear extracts preparation

Cells were collected and incubated with PBS (pH 7.4) containing 0.5% Nonidet-P40 on ice for 30 min, followed by a 10-min centrifugation at 2,000g. The nuclear pellet was collected and frozen at -70 °C. Frozen nuclei were lysed by resuspending in a 100 μl lysis buffer (10 mM Tris-HCl, pH 7.4, 4 mM EDTA, 30 mM KCl, 1% Nonidet-P40 and a 100-fold dilution of protease and phosphatase inhibitor cocktails; Sigma) and incubated on ice for 30 min. The debris was removed by centrifugation at 10,000g for 30 min at 4 °C. The supernatant was collected and the protein amount was determined by Bradford assay.

Heat degradation product of abasic DNA

The heat degradation product from a double-stranded uracil-containing DNA oligonucleotide was prepared as described¹⁹. The sequence of the DNA oligonucleotide is 5'-ATTCCAGAGTGTCAATAACACGGUGGACCAGTCGATCCTGGGCTGCAGGAA TTC-3'.

Received 14 August; accepted 19 November 2001.

- Wilson, D. M., Takeshita, M., Grollman, A. P. & Demple, B. Incision activity of human apurinic endonuclease (Ape) at abasic site analogs in DNA. *J. Biol. Chem.* **270**, 16002–16007 (1995).
- Xanthoudakis, S., Smeys, R. J., Wallace, J. D. & Curran, T. The redox/DNA repair protein, Ref-1, is essential for early embryonic development in mice. *Proc. Natl Acad. Sci. USA* **93**, 8919–8923 (1996).
- Lindahl, T. Suppression of spontaneous mutagenesis in human cells by DNA base excision–repair. *Mutat. Res.* **462**, 129–135 (2000).
- Loeb, L. A. & Preston, B. D. Mutagenesis by apurinic/apyrimidinic sites. *Annu. Rev. Genet.* **20**, 201–230 (1986).
- Matsumoto, Y. & Kim, K. Excision of deoxyribose phosphate residues by DNA polymerase β during DNA repair. *Science* **269**, 699–702 (1995).
- Kunkel, T. A. The mutational specificity of DNA polymerase-β during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations. *J. Biol. Chem.* **260**, 5787–5796 (1985).
- Loeb, K. R. & Loeb, L. A. Significance of multiple mutations in cancer. *Carcinogenesis* **21**, 379–385 (2000).
- Demple, B., Herman, T. & Chen, D. S. Cloning and expression of APE, the cDNA encoding the major human apurinic endonuclease: definition of a family of DNA repair enzymes. *Proc. Natl Acad. Sci. USA* **88**, 11450–11454 (1991).
- Robson, C. N., Milne, A. M., Pappin, D. J. & Hickson, I. D. Isolation of cDNA clones encoding an enzyme from bovine cells that repairs oxidative DNA damage in vitro: homology with bacterial repair enzymes. *Nucleic Acids Res.* **19**, 1087–1092 (1991).
- Chou, K. M., Kukhanova, M. & Cheng, Y. C. A novel action of human apurinic/apyrimidinic endonuclease. Excision of L-configuration deoxyribonucleoside analogs from the 3' termini of DNA. *J. Biol. Chem.* **275**, 31009–31015 (2000).
- Grove, K. L. *et al.* Anticancer activity of β-L-dioxolane-cytidine, a novel nucleoside analogue with the unnatural L configuration. *Cancer Res.* **55**, 3008–3011 (1995).
- Wilson, S. H. & Kunkel, T. A. Passing the baton in base excision repair. *Nature Struct. Biol.* **7**, 176–178 (2000).
- Bennett, R. A., Wilson, D. M., Wong, D. & Demple, B. Interaction of human apurinic endonuclease and DNA polymerase β in the base excision repair pathway. *Proc. Natl Acad. Sci. USA* **94**, 7166–7169 (1997).
- Husain, I. *et al.* Purification and characterization of DNA ligase III from bovine testes. Homology with DNA ligase II and vaccinia DNA ligase. *J. Biol. Chem.* **270**, 9683–9690 (1995).
- Cheng, Y. C., Gao, W. Y., Chen, C. H., Vazquez-Padua, M. & Starnes, M. C. DNA polymerases versus HIV reverse transcriptase in AIDS therapy. *Ann. NY Acad. Sci.* **616**, 217–223 (1990).
- Bouayadi, K. *et al.* Overexpression of DNA polymerase β sensitizes mammalian cells to 2',3'-deoxycytidine and 3'-azido-3'-deoxythymidine. *Cancer Res.* **57**, 110–116 (1997).
- Skalski, V., Liu, S. H. & Cheng, Y. C. Removal of anti-human immunodeficiency virus 2',3'-dideoxynucleoside monophosphates from DNA by a novel human cytosolic 3'-5' exonuclease. *Biochem. Pharmacol.* **50**, 815–821 (1995).

18. Zhu, Q. Y., Scarborough, A., Polsky, B. & Chou, T. C. Drug combinations and effect parameters of zidovudine, stavudine, and nevirapine in standardized drug-sensitive and resistant HIV type 1 strains. *AIDS Res. Hum. Retrovir.* **12**, 507–517 (1996).
19. Strauss, P. R., Beard, W. A., Patterson, T. A. & Wilson, S. H. Substrate binding by human apurinic/apyrimidinic endonuclease indicates a Briggs-Haldane mechanism. *J. Biol. Chem.* **272**, 1302–7. (1997).
20. McCullough, A. K., Dodson, M. L. & Lloyd, R. S. Initiation of base excision repair: glycosylase mechanisms and structures. *Annu. Rev. Biochem.* **68**, 255–285 (1999).
21. Chen, D. S., Herman, T. & Demple, B. Two distinct human DNA diesterases that hydrolyze 3'-blocking deoxyribose fragments from oxidized DNA. *Nucleic Acids Res.* **19**, 5907–5914 (1991).
22. Bhagwat, A. S., Sanderson, R. J. & Lindahl, T. Delayed DNA joining at 3' mismatches by human DNA ligases. *Nucleic Acids Res.* **27**, 4028–4033 (1999).
23. Hoss, M. *et al.* A human DNA editing enzyme homologous to the *Escherichia coli* DnaQ/MutD protein. *EMBO J.* **18**, 3868–3875 (1999).
24. Mazur, D. J. & Perrino, F. W. Identification and expression of the TREX1 and TREX2 cDNA sequences encoding mammalian 3'–5' exonucleases. *J. Biol. Chem.* **274**, 19655–19660 (1999).
25. Mazur, D. J. & Perrino, F. W. Excision of 3' termini by the Trex1 and TREX2 3'–5' exonucleases: characterization of the recombinant proteins. *J. Biol. Chem.* **276**, 17022–17029 (2001).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank Z. Hatahet and J. B. Sweasy for discussion; M. Kelley and B. Demple for providing the APE clones; A. Tomkinson for human DNA ligase I; and J. Sweasy for human DNA polymerase β .

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Y.C.-C. (e-mail: Cheng.lab@Yale.edu).

Mechanism of force generation by myosin heads in skeletal muscle

Gabriella Piazzesi*, Massimo Reconditi*, Marco Linari*, Leonardo Lucii*, Yin-Biao Sun†, Theyencheri Narayanan‡, Peter Boesecke‡, Vincenzo Lombardi* & Malcolm Irving†

* Università di Firenze, Viale G.B. Morgagni 63, I-50134 Firenze, Italy

† Randall Centre, School of Biomedical Sciences, King's College London, London SE1 1UL, UK

‡ European Synchrotron Radiation Facility, F-38043 Grenoble Cedex, France

Muscles generate force and shortening in a cyclical interaction between the myosin head domains projecting from the myosin filaments and the adjacent actin filaments. Although many features of the dynamic performance of muscle are determined by the rates of attachment and detachment of myosin and actin¹, the primary event in force generation is thought to be a conformational change or 'working stroke' in the actin-bound myosin head^{2–8}. According to this hypothesis, the working stroke is much faster than attachment or detachment, but can be observed directly in the rapid force transients that follow step displacement of the filaments³. Although many studies of the mechanism of muscle contraction^{9–13} have been based on this hypothesis, the alternative view—that the fast force transients are caused by fast components of attachment and detachment^{14–17}—has not been excluded definitively. Here we show that measurements of the axial motions of the myosin heads at ångström resolution by a new X-ray interference technique¹⁸ rule out the rapid attachment/detachment hypothesis, and provide compelling support for the working stroke model of force generation.

When the length of an active muscle fibre is decreased suddenly, so that each set of myosin filaments slides by a few nanometres along

the neighbouring actin filaments, the force decreases during the length change (Fig. 1a). This reflects the undamped elasticity of the muscle fibre, and is called phase 1 of the force transient^{3,19}. After the length change, force recovers at about $1,000 \text{ s}^{-1}$ (phase 2). Our aim was to determine whether the phase 2 force recovery is caused by a working stroke in the actin-attached myosin heads^{2–8}, or by rapid detachment of heads followed by rapid attachment to different actin monomers^{14–17}.

We measured the axial motion of the myosin heads during the force transients using X-ray interference between the two arrays of myosin heads in each filament (Fig. 1c). Each half of the myosin filament (magenta) contains an array of 49 layers of heads (orange) with a regular spacing d ($\sim 14.5 \text{ nm}$). These arrays give rise to an axial X-ray reflection called the M3, with intensity distribution $\sin^2(49\pi R d)/\sin^2(\pi R d)$ (Fig. 1d, orange), where R is the reciprocal space parameter. The two arrays are separated by a 'bare zone' of length B ($\sim 160 \text{ nm}$), so that their centres are a distance $L = B + 48d$ apart (Fig. 1c). X-ray interference between the two arrays effectively multiplies the intensity distribution by $\cos^2(\pi R L)$ (Fig. 1d, magenta). The resulting fine structure of the X-ray reflection (Fig. 1d, green) provides an extremely sensitive measure of the axial motion of the myosin heads¹⁸.

We recorded the intensity profile of the M3 reflection in a 2-ms period before the length step (Fig. 1a, b, T_0 , green), in a 100- μs period close to the end of phase 1 (T_1 , red), and in a 2-ms period near the end of phase 2 (T_2 , blue). The length step caused a decrease in the relative intensity of the higher angle peak of the M3 reflection, and shifted both peaks to a higher angle (Fig. 1b). These changes are in the direction expected for a decrease of the interference distance (L) between the two arrays of actin-attached myosin heads when the fibre shortens and the actin filaments move towards the centre of the myosin filament (Fig. 1c). Most of the change in L occurred during phase 1 of the force transient (Fig. 1b, green to red).

The changes in interference fine structure of the M3 reflection were expressed as the ratio of the intensity of the higher angle peak

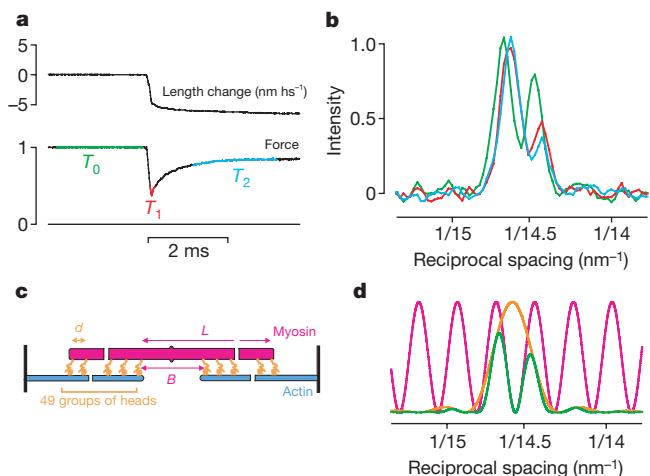


Figure 1 Changes in interference fine structure of the M3 X-ray reflection produced by rapid shortening. **a**, Length change in nanometres per half-sarcomere (hs), and force normalized by isometric force T_0 . Fibre cross-sectional area, $27,300 \mu\text{m}^2$; T_0 , 293 kN m^{-2} ; sarcomere length, $2.13 \mu\text{m}$. **b**, Axial X-ray intensity distribution of the M3 reflection normalized to that of the lower angle peak. Colours denote the periods T_0 , T_1 and T_2 shown in **a**; the total exposure times (equivalent unattenuated beam) were 52, 158, and 105 ms, respectively. **c**, The two arrays of myosin heads (orange) in each myosin filament (magenta). **d**, Axial X-ray intensity distribution (green) calculated as the product of the diffracted intensity from an array of 49 heads with $d = 14.57 \text{ nm}$ (orange) and the interference function for $L = 865.92 \text{ nm}$ (magenta).

to that of the lower angle peak, I_{HA}/I_{LA} (Fig. 2a), and the spacings of the lower and higher angle peaks, S_{LA} and S_{HA} (Fig. 2b). All three parameters decreased as the size of the shortening step was increased, although the changes seemed to saturate for steps larger than 7 nm. The observed changes were similar at T_1 (red circles) and T_2 (blue triangles). The mean spacing of the M3 reflection, S_{M3} (Fig. 2b), which is related to the elastic strain of the myosin filament, was less steeply dependent on the size of the applied length step.

We used these experimental values of I_{HA}/I_{LA} and S_{M3} to calculate the interference distance L and the myosin head spacing d using the theoretical intensity distributions described above (Fig. 1c). These equations would only apply exactly if each layer of myosin heads could be considered as a point mass, but a similar analysis using the axial mass projections of atomic models for the myosin head^{5–7} gave the same values of I_{HA}/I_{LA} and S_{M3} within the precision of the present measurements¹⁸. The changes in L can therefore be interpreted solely in terms of the average axial motion (Δc) of the centres of mass of the myosin heads with respect to their attachments to the

myosin filament, independent of any other effects of changes in their conformation.

The X-ray interference method allows Δc to be measured with a precision of about 1 Å (Fig. 3a). Δc was roughly proportional to the extent of the shortening step for steps of up to 7 nm per half-sarcomere, but was about five times smaller than the imposed filament sliding. The value of Δc at T_2 (Fig. 3a, blue triangles) was similar to that at T_1 (red circles). This result is consistent with models in which force is generated in phase 2 by a working stroke in actin-attached myosin heads, but is not consistent with models that lack a working stroke.

In both types of model Δc is expected to decrease in phase 1, because actin-attached myosin heads are carried towards the centre of the myosin filament during muscle shortening. In working stroke models, the heads remain attached to actin in phase 2; there is no filament sliding in this period and so, neglecting for the moment any effects of filament compliance, the actin-binding sites of the heads would remain at the same distance from the centre of the myosin filament. In contrast, in models without a working stroke force is simply proportional to the strain in the head; thus, force generation in phase 2 would be accompanied by myosin heads binding to new actin sites that were, on average, farther from the centre of the myosin filament. This would produce an increase in Δc during phase 2, which was not observed (Fig. 3a).

To make a quantitative interpretation of the observed axial

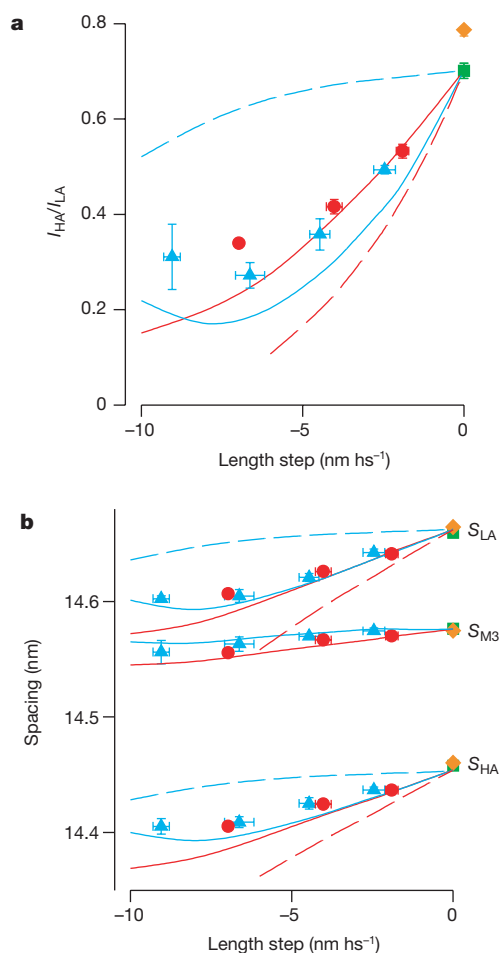


Figure 2 Relative intensity and spacings of the higher (HA) and lower (LA) angle peaks of the M3 reflection. **a**, Relative intensity; **b**, spacings. S_{M3} , weighted mean spacing. Red circles (T_1) and blue triangles (T_2) are the means $\pm$ s.e.m. for $n = 3$ –6 fibres, except for the T_1 point at -7 nm per half-sarcomere ($n = 1$). Green squares denote values for the 2-ms period before the length step (T_0 , Fig. 1a); orange diamonds are values from contractions without imposed length steps. The lines were calculated from the simulations described in the text: red, T_1 ; blue, T_2 . Continuous lines indicate that only one head of each myosin responds to the length step; red dashed line (T_1) indicates that all heads in the fibre respond to the step; blue dashed line (T_2) indicates the rapid detachment/attachment model.

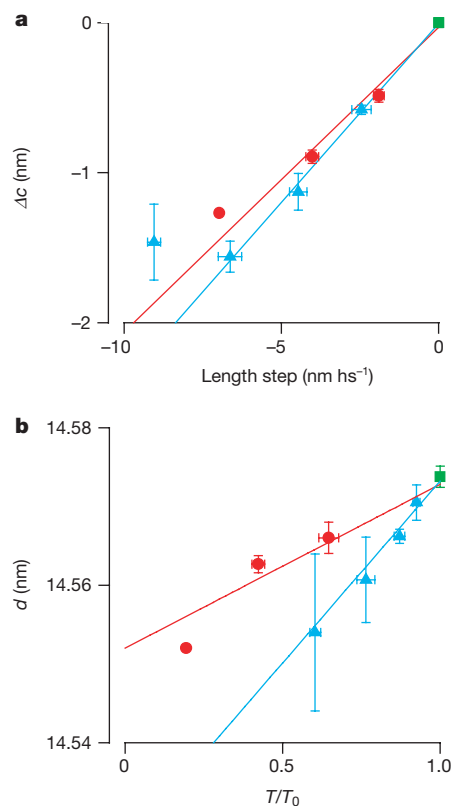


Figure 3 Axial motions and spacings of myosin heads. **a**, Axial motions (Δc) of myosin head centroids with respect to their myosin filament attachments. **b**, Axial spacing of the heads along the myosin filament (d) plotted against normalized force T/T_0 . Parameters Δc and d were obtained for each fibre and condition by fitting experimental values of I_{HA}/I_{LA} and S_{M3} (Fig. 2) for each fibre. A negative value of Δc denotes motion towards the centre of the myosin filament. See Fig. 2 for definition of symbols. The lines were obtained by linear regression of all individual fibre data except the T_2 points in the -9 nm per half-sarcomere class in **a**.

motions of the myosin heads (Δc), the compliance of the actin and myosin filaments must be taken into account. The filament compliances are usually expressed as the average strain in the actin or myosin filament, $\langle S_A \rangle$ or $\langle S_M \rangle$, owing to the active isometric force (T_0). $\langle S_A \rangle$ is 0.26%/T₀ in the conditions used here^{20,21}. $\langle S_M \rangle$ is the slope of the relationship between the myosin filament periodicity d and the normalized force T/T_0 . $\langle S_M \rangle$ was $0.14 \pm 0.03\%/T_0$ at T_1 (Fig. 3b, red circles, $n = 9$ fibres) and $0.34 \pm 0.05\%/T_0$ at T_2 (blue triangles, $n = 18$).

The value of $\langle S_M \rangle$ at T_1 represents the instantaneous compliance of the myosin filament, and the higher value at T_2 is likely to include a contribution from changes in the structure of the myosin filament^{18,22}. The total instantaneous compliance of each set of overlapping myosin and actin filaments crosslinked by myosin heads during active isometric contraction is 5.1 nm/T₀. Of this, roughly 40% is due to the compliance of the actin filaments, 20% to that of the myosin filaments, and 40% to that of the myosin heads^{20,21}.

We calculated the axial displacement of each layer of myosin heads along the filament during the elastic phase 1 response to a length step from these filament and myosin head compliances. These motions occur to the same extent in models with and without a working stroke, so the phase 1 analysis provides an independent calibration of the X-ray interference method. Each layer of heads was represented by an atomic model with the catalytic domain in its nucleotide-free conformation^{5,6} and a variable bend between the catalytic and light chain domains^{7,8} (see Methods). The changes in I_{HA}/I_{LA} , S_{LA} and S_{HA} calculated from this model (Fig. 2, red dashed lines) were much larger than the observed changes (red circles).

The discrepancy is too large to be explained by uncertainty about the location of the pivot in the myosin head; it was reduced by only about a quarter when the pivot was assumed to be at the actin-binding site. Moreover, the discrepancy cannot be explained by possible errors in the values of filament compliance. The calculated values of I_{HA}/I_{LA} , S_{LA} and S_{HA} are almost independent of the value of $\langle S_M \rangle$, because myosin filament compliance leads to proportional changes in d and L (ref. 18). The discrepancy is reduced by increasing the value of $\langle S_A \rangle$, but would not be eliminated even if $\langle S_A \rangle$ were twice the measured value^{20,21}, in which case almost the whole compliance of the muscle sarcomere would be in the filaments, with virtually no contribution left for the myosin heads.

We conclude that this type of model, in which the M3 reflection is solely due to the myosin heads, and all the myosin heads in the fibre respond to filament sliding, is incompatible with the observed changes in the fine structure of the M3 reflection. It is unlikely that other structures in the fibre make a substantial contribution to the active M3 reflection, because the reflection can be almost completely abolished by extended high-velocity shortening²². Moreover, the linear dependence of the intensity of the active M3 reflection on the degree of overlap between the myosin and actin filaments¹⁸ shows that it originates from the part of the myosin filament that overlaps with actin, and that detached myosin heads in the region of the filament that does not overlap with actin are too disordered to make a significant contribution.

Thus, the smaller-than-expected changes in interference fine structure observed during length steps show that there is a population of myosin heads in the overlap region that do not respond to filament sliding, although they have sufficient axial order to contribute to the M3 reflection. Because each myosin molecule has two heads, these static heads are likely to be the partners of heads that are attached to actin. The two heads are joined at the head–tail junction, so attachment of one head to actin would impose axial order on the other head, and both would contribute to the M3 X-ray reflection. Adding the partner heads to the model gave a reasonably good fit to the observed changes in

I_{HA}/I_{LA} , S_{LA} and S_{HA} at T_1 (Fig. 2, red continuous lines), especially considering that there are no adjustable parameters in the model. Thus the elastic, phase 1 changes in the interference fine structure can be explained quantitatively if only one head of each myosin molecule bears the force of isometric contraction and responds to filament sliding.

Finally, we return to the quantitative explanation of the origin of rapid force recovery in phase 2 of the force transient. If, as argued above, only one head of each myosin responds to the length step, would a working stroke in this head reproduce the observed values of I_{HA}/I_{LA} , S_{LA} and S_{HA} at T_2 ? The calculated values for T_2 for the working stroke model (Fig. 2, blue continuous lines) gave a good but not perfect fit to the experimental data (blue triangles). In contrast, the rapid attachment/detachment model (Fig. 2, blue dashed lines) was clearly inconsistent with the observed changes, independent of the magnitude of filament compliance. The imperfect fit of the working stroke model at T_2 is probably caused by a small fraction of heads detaching during phase 2; both stiffness measurements¹¹ and changes in the intensity of the M3 reflection produced by closely spaced pairs of length steps¹² suggest that about 8% of the heads have detached by T_2 after a 5-nm shortening step and have become disordered axially, so that they no longer contribute to the M3 reflection or to the axial displacement Δc measured from its interference fine structure.

Most of the heads behave as predicted by the working stroke model. This model reproduces the observed nonlinear dependence of the interference parameters on the size of length step, and the approximate fit was maintained for the largest releases studied (9 nm). Thus, our results provide strong support for working stroke models in which myosin heads remain attached during filament sliding of about 10 nm, and are clearly inconsistent with models in which force is generated by rapid attachment and detachment of myosin and actin. □

Methods

Mechanical measurements

Single fibres from the anterior tibialis muscle of *Rana temporaria* were mounted horizontally between a loudspeaker coil motor and a capacitance gauge force transducer²³ at sarcomere length 2.1 μm , 4 °C, and electrically stimulated for 2.3 s at 18–25 Hz. After 0.3 s of isometric contraction, forty 50-ms shortening/stretch cycles were imposed with a 4-ms interval between shortening and stretch¹². We measured sarcomere length continuously in a segment of 1–2 mm near the fibre centre with a striation follower²⁴.

X-ray data collection

The fibre and mechanical apparatus were mounted vertically on X-ray beamline ID2A of the European Synchrotron Radiation Facility (ESRF), Grenoble, France²⁵. The X-ray beam had a full width at half maximum (FWHM) of about 0.1 mm vertically and 0.6 mm horizontally, a flux of up to 2×10^{13} photons s^{−1} and a wavelength of 0.1 nm. We controlled X-ray exposure with 10- μs precision by two electromagnetic shutters in series. Diffraction patterns were recorded on A3-size storage PhosphorImage plates (Molecular Dynamics) mounted in an evacuated camera tube 9.85 m from the fibre. Image plates were scanned off-line (Molecular Dynamics 840 scanner) with 100- μm pixels. The vertical point spread function of the X-ray camera/scanner combination had FWHM 320 μm .

Experimental protocol

We stimulated fibres at 4-min intervals and moved them vertically by 100–250 μm between contractions to spread the effects of radiation damage. Diffraction data for 100- μs exposures were typically accumulated from 40 length steps in each of 20 contractions with the unattenuated X-ray beam (total exposure 80 ms). For 2-ms exposures, we used 40 length steps, 4 contractions and 4 $\times$ beam attenuation. Diffraction data during isometric contraction were recorded with 200-ms exposures, 4 contractions and 7 $\times$ beam attenuation. Total X-ray exposure per fibre was typically equivalent to 0.5 s of unattenuated beam. Experiments were terminated by failure of excitation–contraction coupling; up to this point X-ray and mechanical responses were stable. Data are presented from 18 fibres with a cross-sectional area of $22,500 \pm 5,800 \mu\text{m}^2$ (mean $\pm$ s.d.).

X-ray data analysis

We analysed X-ray diffraction patterns with the programs HV (A. Stewart, Brandeis), Fit2D (A. Hammersley, ESRF) and Peakfit (SPSS Science). Patterns were aligned and centred using the 1,0 equatorial reflections. The background under the axial X-ray

reflections was subtracted using HV. The axial intensity distribution was calculated by radial integration from $\pm 1/64 \text{ nm}^{-1}$ and the M3 reflection fitted by two gaussian peaks using Peakfit. We calculated the M3 spacing (S_{M3}) as the intensity-weighted mean of that of the two peaks. Spacings were calibrated assuming that S_{M3} is 14.340 nm at rest²⁶ or 14.573 nm during active isometric contraction¹⁸.

Simulation of the axial intensity distribution

The interference fine structure of the M3 reflection was calculated as the Fourier transform of the axial mass distribution. Each myosin filament was assumed to contain two arrays of myosin heads separated by a bare zone^{27,28} of length B ($\sim 160 \text{ nm}$). Each array has 49 layers of heads with axial periodicity d . The axial mass projection of each layer of heads was calculated from crystallographic data, with the catalytic domain (CD, heavy chain residues 1–707) of the head in its nucleotide-free actin-bound conformation^{5,6}, and the light chain domain (LCD, heavy chain residues 707–843 and both light chains) tilted axially around the CD/LCD junction^{7,8,12}. During isometric contraction the 707–843 axis was assumed to be at 60° to the filament axis¹². LCD tilt dispersion of $\pm 20^\circ$, which would allow myosin heads to bind to actin monomers with axial periodicity of 5.5 nm , had a negligible effect on the results¹².

We used a distributed filament compliance formalism^{21,29,30} to calculate the force and strain distribution along the myosin and actin filaments. The strain in myosin heads during isometric contraction was assumed to be uniform along the filament overlap zone. Rapid force recovery after a length step was assumed to be due to a uniform force contribution from heads along the overlap zone. We calculated the mass distribution along the filament with the position of the head–rod junction of each level of heads determined by the myosin filament strain; head strain was represented as a tilt of the LCD with respect to the actin-bound CD⁵. In the working stroke model, the distribution of myosin head strain along the overlap zone after a length step was calculated under the constraint of constant sarcomere length during rapid force recovery. Force recovery during the $110\text{-}\mu\text{s}$ length step itself was estimated from the difference between the measured force (T_1) and that expected from the instantaneous half-sarcomere compliance ($5.1 \text{ nm}/T_0$; ref. 20) and taken into account in calculating the mass distribution at T_1 . Simulations using the rapid detachment/attachment model assumed that the total fraction of myosin heads attached to actin remained constant and that the average head strain was proportional to the force.

Received 24 July; accepted 3 December 2001.

- Huxley, A. F. Muscle structure and theories of contraction. *Prog. Biophys. Biophys. Chem.* **7**, 255–318 (1957).
- Huxley, H. E. The mechanism of muscle contraction. *Science* **164**, 1356–1366 (1969).
- Huxley, A. F. & Simmons, R. M. Proposed mechanism of force generation in striated muscle. *Nature* **233**, 533–538 (1971).
- Huxley, A. F. Muscular contraction (review lecture). *J. Physiol. (Lond.)* **243**, 1–43 (1974).
- Rayment, I. *et al.* Three-dimensional structure of myosin subfragment-1: a molecular motor. *Science* **261**, 50–58 (1993).
- Rayment, I. *et al.* Structure of the actin–myosin complex and its implications for muscle contraction. *Science* **261**, 58–65 (1993).
- Dominguez, R., Freyzon, Y., Trybus, K. M. & Cohen, C. Crystal structure of a vertebrate smooth muscle myosin motor domain and its complex with the essential light chain: visualisation of the pre-power stroke state. *Cell* **94**, 559–571 (1998).
- Gees, M. A. & Holmes, K. C. Structural mechanism of muscle contraction. *Annu. Rev. Biochem.* **68**, 687–728 (1999).
- Huxley, H. E. *et al.* Changes in the X-ray reflections from contracting muscle during rapid mechanical transients and their structural implications. *J. Mol. Biol.* **169**, 469–506 (1983).
- Irving, M., Lombardi, V., Piazzesi, G. & Ferenczi, M. A. Myosin head movements are synchronous with the elementary force-generating process in muscle. *Nature* **357**, 156–158 (1992).
- Lombardi, V., Piazzesi, G. & Linari, M. Rapid regeneration of the actin–myosin power stroke in contracting muscle. *Nature* **355**, 638–641 (1992).
- Irving, M. *et al.* Conformation of the myosin motor during force generation in skeletal muscle. *Nature Struct. Biol.* **7**, 482–485 (2000).
- Corrie, J. E. T. *et al.* Dynamic measurement of myosin light-chain domain tilt and twist in muscle contraction. *Nature* **400**, 425–430 (1999).
- Podolsky, R. J. & Nolan, A. C. Muscle contraction transients, cross-bridge kinetics, and the Fenn effect. *Cold Spring Harb. Symp. Quant. Biol.* **37**, 661–668 (1973).
- Brenner, B. Rapid dissociation and reassociation of actomyosin cross-bridges during force generation: a newly observed facet of cross-bridge action in muscle. *Proc. Natl Acad. Sci. USA* **88**, 10490–10494 (1991).
- Kitamura, K., Tokunaga, M., Iwane, A. H. & Yanagida, T. A single myosin head moves along an actin filament with regular steps of 5.3 nm . *Nature* **397**, 129–134 (1999).
- Howard, J. *Mechanics of Motor Proteins and the Cytoskeleton* (Sinauer, Sunderland, MA, 2001).
- Linari, M. *et al.* Interference fine structure and sarcomere length dependence of the axial X-ray pattern from active single muscle fibres. *Proc. Natl Acad. Sci. USA* **97**, 7226–7231 (2000).
- Ford, L. E., Huxley, A. F. & Simmons, R. M. Tension responses to sudden length change in stimulated frog muscle fibres near slack length. *J. Physiol. (Lond.)* **269**, 441–515 (1977).
- Dobbie, I. *et al.* Elastic bending and axial tilting of myosin heads during muscle contraction. *Nature* **396**, 383–387 (1998).
- Linari, M. *et al.* The stiffness of skeletal muscle in isometric contraction and rigor: the fraction of myosin heads bound to actin. *Biophys. J.* **74**, 2459–2473 (1998).
- Piazzesi, G. *et al.* Changes in conformation of myosin heads during the development of isometric contraction and rapid shortening in single frog muscle fibres. *J. Physiol. (Lond.)* **514**, 305–312 (1999).
- Lombardi, V. & Piazzesi, G. The contractile response during steady lengthening of stimulated frog muscle fibres. *J. Physiol. (Lond.)* **431**, 141–171 (1990).

- Huxley, A. F., Lombardi, V. & Peachey, L. D. A system for fast recording of longitudinal displacement of a striated muscle fibre. *J. Physiol. (Lond.)* **317**, 12P–13P (1981).
- Boescke, P., Diat, O. & Rasmussen, B. High-brilliance beamline at the European Synchrotron Radiation Facility. *Rev. Sci. Instrum.* **66**, 1636–1638 (1995).
- Haselgrove, J. C. X-ray evidence for conformational changes in the myosin filaments of vertebrate striated muscle. *J. Mol. Biol.* **92**, 113–143 (1975).
- Sjöström, M. & Squire, J. M. Fine structure of the A band in cryosections: the structure of the A-band of human skeletal muscle fibres from ultrathin cryosections negatively stained. *J. Mol. Biol.* **109**, 49–68 (1977).
- Craig, R. Structure of A-segments from frog and rabbit skeletal muscle. *J. Mol. Biol.* **109**, 69–81 (1977).
- Thorson, J. W. & White, D. C. S. Distributed representations for actin-myosin interaction in the oscillatory contraction of muscle. *Biophys. J.* **9**, 360–390 (1969).
- Ford, L. E., Huxley, A. F. & Simmons, R. M. The relation between stiffness and filament overlap in stimulated frog muscle fibres. *J. Physiol. (Lond.)* **311**, 219–249 (1981).

Acknowledgements

This work was supported by the Medical Research Council, Consiglio Nazionale delle Ricerche (CNR), Ministero dell'Istruzione, dell'Università e della Ricerca (MURST), Telethon (Italy), EMBL, EU and ESRF. We thank A. F. Huxley for comments, J. Gorini, A. Aiazzi and M. Dolfi for mechanical and electronics support.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.I. (e-mail: malcolm.irving@kcl.ac.uk).

RanGAP mediates GTP hydrolysis without an arginine finger

Michael J. Seewald, Carolin Körner, Alfred Wittinghofer & Ingrid R. Vetter

Max-Planck-Institut für molekulare Physiologie, Abteilung Strukturelle Biologie, Postfach 102664, D-44026 Dortmund, Germany

GTPase-activating proteins (GAPs) increase the rate of GTP hydrolysis on guanine nucleotide-binding proteins by many orders of magnitude. Studies with Ras and Rho have elucidated the mechanism of GAP action by showing that their catalytic machinery is both stabilized by GAP binding and complemented by the insertion of a so-called 'arginine finger' into the phosphate-binding pocket^{1,2}. This has been proposed as a universal mechanism for GAP-mediated GTP hydrolysis. Ran is a nuclear Ras-related protein that regulates both transport between the nucleus and cytoplasm during interphase, and formation of the mitotic spindle and/or nuclear envelope in dividing cells³. Ran–GTP is hydrolysed by the combined action of Ran-binding proteins (RanBPs) and RanGAP⁴. Here we present the three-dimensional structure of a Ran–RanBP1–RanGAP ternary complex in the ground state and in a transition-state mimic. The structure and biochemical experiments show that RanGAP does not act through an arginine finger, that the basic machinery for fast GTP hydrolysis is provided exclusively by Ran and that correct positioning of the catalytic glutamine is essential for catalysis.

Ran-mediated GTP hydrolysis is thought to be the main driving force behind cargo transport across the nuclear pore complex. The location of the Ran-specific guanine nucleotide exchange factor RCC1 in the nucleus and that of RanGAP in the cytoplasm creates a gradient of Ran–GTP across the nuclear pore complex. The GTP-bound form of Ran in the nucleus binds with high affinity to the import receptors of the importin- β family and thereby induces the release of import cargo, whereas export cargo binding to export receptors (exportins) requires the presence of Ran–GTP³. The

dissociation of Ran–GTP complexes with importins and exportins on the cytoplasmic side of the nuclear pore complex requires the action of the Ran-binding protein RanBP1 or RanBP2 (ref. 5) and is made irreversible by RanGAP-mediated GTP hydrolysis.

RanGAP was first isolated biochemically before it was found to constitute a homologue of a *Saccharomyces cerevisiae* mutant that is defective in RNA processing or transport⁶. The structure of Rna1, the yeast homologue of RanGAP (hereafter referred to as RanGAP), has been solved⁷ and shown to contain leucine-rich repeats (LRRs). Kinetic studies showed that Arg 74 is important for catalysis, which supported the idea that RanGAP may act through a canonical arginine finger comparable to that of rasGAP¹ or rhoGAP². The structure of a ternary Ran–RanBP1–RanGAP complex presented here shows, however, that RanBP1 does not participate directly in mediating hydrolysis and that the GTPase mechanism does not involve an arginine finger.

We solved the structure of the ternary complex (Fig. 1) by

molecular replacement using free RanGAP⁷ as a search model. Ran sits on top of the crescent-shaped arch of RanGAP, with RanBP1 lying on top of Ran without contacting RanGAP (Fig. 1). The binding site of RanGAP corresponds, in principle, to a prediction made on the basis of sequence conservation and mutagenesis studies⁷, although the details were not previously anticipated. Superimposing the ternary complex with the Ran–importin- β and Ran–transportin complexes^{8,9} on top of Ran seems to support the biochemical evidence that the binding of import receptors and RanGAP to Ran is mutually exclusive.

Interactions between Ran and RanGAP (Figs 1b and 2a) are mediated by 7 of the 11 LRRs. The protruding loop in LRR3 seems to act as a 'footrest' for Ran and stabilizes the switch II region of Ran, which, in analogy to other guanine nucleotide-binding proteins (GNBPs), may be flexible in uncomplexed Ran–GTP. The interface between Ran and RanGAP, which buries 3,616 Å² of solvent-accessible surface, is dominated largely by strong charge–charge interactions. The acidic carboxy terminus of RanGAP, which is important for binding affinity¹⁰, is not completely visible, as in uncomplexed RanGAP. But it is in close vicinity to the basic patch in Ran (residues 139–142).

The structure shows that Leu 43 of Ran, mutation of which makes

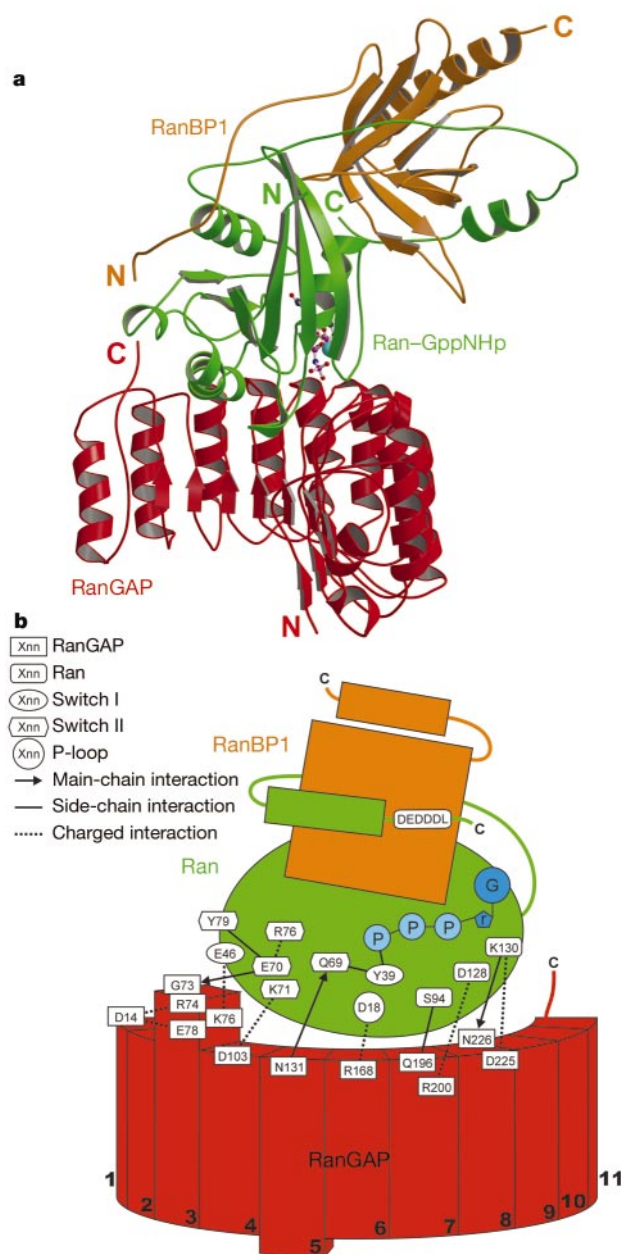


Figure 1 Structure of the ternary complex. **a**, Ribbon plot of the overall structure of the Ran–RanBP1–RanGAP complex. **b**, Overview of interaction between Ran and RanGAP.

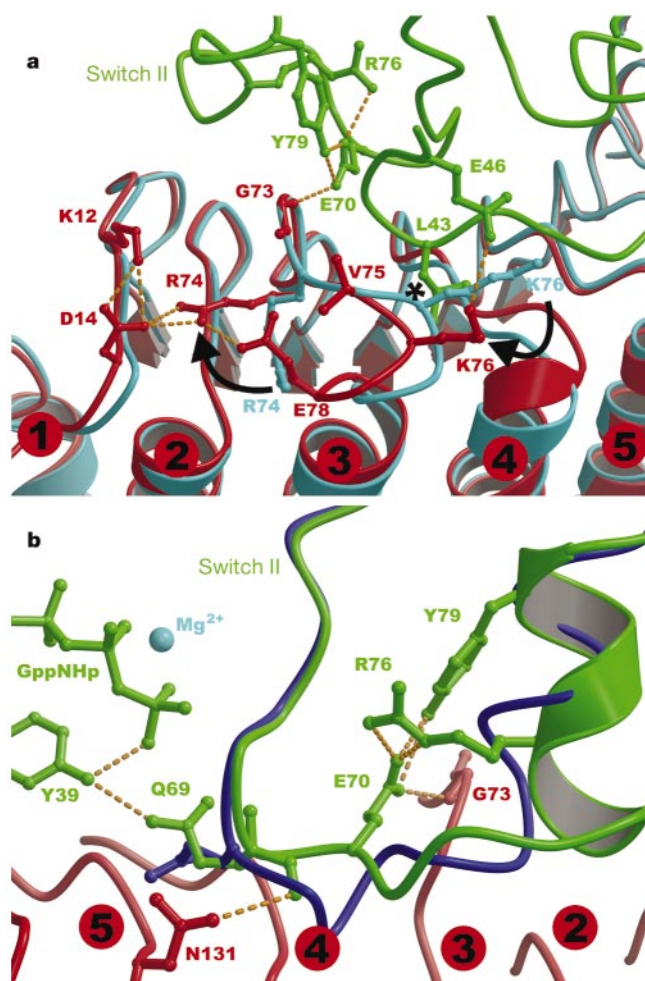


Figure 2 Details of the Ran–RanGAP interface. **a**, Worm plot of Ran and RanGAP, with colours as in Fig. 1 and important residues shown in ball and stick representation. The uncomplexed RanGAP (cyan) is superimposed; arrows indicate the movement of Lys 76 and Arg 74, and asterisk symbolizes the potential clash of Leu 43 and Lys 76 on complex formation. **b**, Switch II of Ran (green) superimposed on the Ran–RanBP1 structure (blue), illustrating the reorientation of Gln 69 into a catalytically competent conformation by the residues Tyr 39 from Ran and Asn 131 from RanGAP.

Ran insensitive to stimulation by RanGAP¹¹, is located close to the footrest (Fig. 2a). This residue is probably responsible for the structural change in RanGAP on complex formation, as it would clash with Lys 76 in free RanGAP. Several mutations have been introduced into RanGAP that show that the salt bridge between Asp 225 of RanGAP and Lys 130 of Ran (Fig. 2b) contributes to the overall reactivity, whereas that between Arg 168 of RanGAP and Asp 18 of Ran (Fig. 2b) does not. There are no interactions between Ran and Lys 9, Lys 12 and Glu 78 of RanGAP, which confirms previous biochemical data⁷. In RanGAP, Arg 74, which affects catalysis strongly when mutated, does not interact with Ran nor with the nucleotide (see below and Fig. 2a).

Although RanBP1 is a coactivator of RanGAP-mediated hydrolysis⁴, the intrinsic GTPase reaction is only marginally accelerated by RanBP1 (Supplementary Information). The overall structures of the binary Ran–RanBD1 complex (ref. 12) and our ternary complex

with RanBP1 are very similar, except for minor alterations that are caused mostly by sequence changes. In addition, more residues from the amino terminus and the C-terminal D-E-D-D-D-L amino-acid motif of Ran are visible in the ternary complex. From the structure of the ternary complex, it seems unlikely that RanBP1 modifies the properties of RanGAP. This raises the issue of how coactivation is achieved. The rate of RanGAP-mediated GTP hydrolysis is accelerated in the absence of the C-terminal D-E-D-D-D-L motif and co-stimulation of the reaction by RanBP1 is dependent on the presence of this motif¹³. We conclude that the RanGAP interaction with Ran

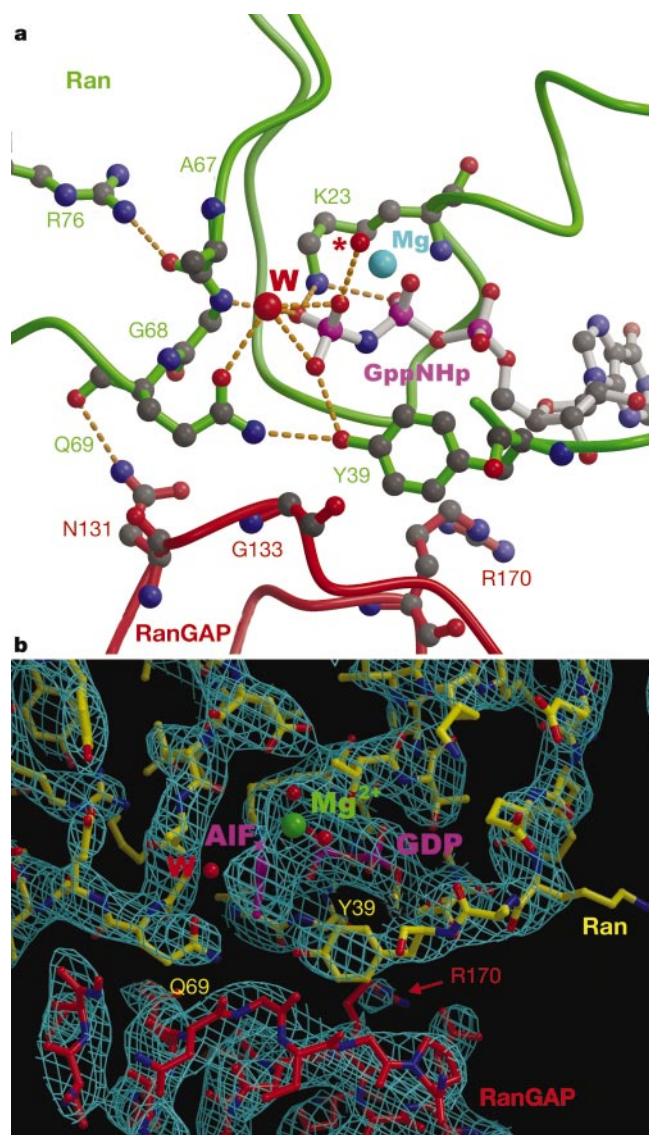


Figure 3 The active site. **a**, The nucleotide and relevant residues of Ran and RanGAP; interactions are indicated by dashed lines. The catalytic water ('W') has very weak density (probably owing to the limited resolution) and was not included in the final model, but is shown here to illustrate its potential interaction partners. It sits in a similar position as in the Ras–RasGAP complex. Asterisk denotes the hydroxyl group of Thr 42 of Ran. The rest of the side chain is omitted for clarity. **b**, The $2F_o - F_c$ electron density map contoured at 1.6σ for the structure with GDP and aluminium fluoride, the latter modelled as AlF_3 .

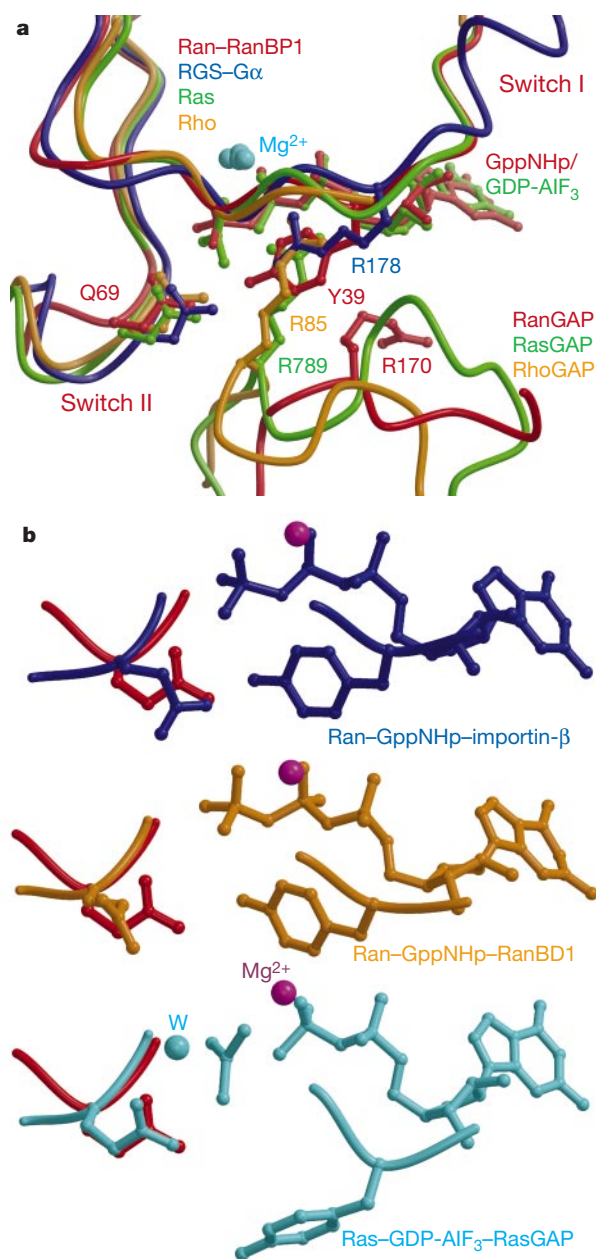


Figure 4 Position of the catalytic glutamine. **a**, Superimposition of the Ran–RanBP1–RanGAP (red), Rho–rhoGAP (orange), Ras–rasGAP (green) and $\text{G}\alpha$ –RGS (blue) complexes, together with the nucleotides, Mg^{2+} and the AlF_3 from the Ras–RasGAP complex, as indicated. Note the very similar architecture of the switch I and II loops, the similar location of the glutamine residues, and the clustering of *cis*-arginine from $\text{G}\alpha$, *trans*-arginine from Ras and RhoGAPs, and the tyrosine residue from Ran (Tyr 32 from Ras and Cdc42 were left out for clarity). **b**, Superimposition of the catalytic Gln 69 of the Ran–RanBP1–RanGAP complex (red) with Ran–GppNHp–importin- β (dark blue), Ran–GppNHp–RanBD1 (orange) and Gln 61 from Ras–RasGAP (cyan).

is inhibited by the C terminus of Ran, and that this inhibition is relieved by RanBP1.

In the complex, the structure of RanGAP is essentially unchanged compared to that of free RanGAP before, the C-terminal stretch of 41 residues is unstructured. In the complex it is located close to the basic patch on Ran that is occupied by the D-E-D-D-D-L motif in Ran-GDP¹⁴. Differences are observed in the protruding loop of LRR3 (residues 71–78), the most highly conserved sequence element of RanGAP (Fig. 2a). The active site in the ternary complex (Fig. 3a) shows well defined electron density for all side chains. Most of the interactions between nucleotide and protein involve canonical interactions found in other GNBPs, such that Mg²⁺ and the phosphate loop (P-loop) lysine (Lys 23) of Ran are both contacting β - and γ -phosphate oxygens, and Thr 42 and Gly 68 of Ran interact with γ -phosphate oxygens.

Unexpectedly, there is no arginine contacting the γ -phosphate oxygens or the β - γ -bridging phosphate oxygen, as in other GAP complexes. Although theoretically the side chain of Arg 170 of RanGAP could reach into the active site, it is bent away from the γ -phosphate (Fig. 3a) and forms a salt bridge with Glu 172 of RanGAP that is not crucial for catalysis^{7,15}. Arginine 74 of RanGAP, which had been a strong suspect for a catalytic arginine⁷, is located too far away (17.5 Å between the γ -phosphate and the ϵ -nitrogen). Its function seems to be purely structural, as it forms salt bridges with Asp 14 and Glu 78 of RanGAP that stabilize the footrest of RanGAP (Fig. 2a).

Instead of an arginine, Tyr 39 of Ran forms hydrogen bonds with both the γ -phosphate oxygen and the Gln 69 side chain of Ran (Fig. 3a). This tyrosine is in the same position as the homologous Tyr 32 of Cdc42 in a complex with Cdc42GAP in which the catalytic arginine has been mutated and its position taken over by Tyr 32 (ref. 16). To probe the importance of Tyr 39 in Ran, we analysed the properties of a Tyr39Ala mutant, which turned out to have a 3-fold weaker affinity and a 4.6-fold lower catalytic rate constant (k_{cat}) than the wild type (Supplementary Information). We could not analyse the properties of a Tyr39Phe mutant because it was unstable, but from the Tyr39Ala mutant it follows that removing the aromatic side chain reduces the value of k_{cat}/K_m (Michaelis constant) 14-fold.

Because studies on rhoGAP have shown that large reorientations occur between the ground state and the transition-state mimic represented by GDP and aluminium fluoride², we also solved the structure of the corresponding aluminium fluoride complex with GDP. Although the resolution of this data set is only 3.1 Å, the quality of the electron density (Fig. 3b) is good enough to show the aluminium fluoride in the position of the γ -phosphate and the surrounding side chains. Gel filtration followed by atomic absorption spectroscopy proved the presence of aluminium fluoride in the complex. Compared with the ground-state complex with GppNHp, the structure of the transition-state mimic showed no significant differences in the position of the amino-acid side chains.

Superimposing the structures for Ran-RanBP1-RanGAP, Ras-rasGAP¹, Rho-rhoGAP² and G α -RGS (regulator of G-protein signalling)¹⁷ shows that features such as the P-loop, switch I and switch II (Fig. 4a) are conserved among the active sites. The only big difference between Ran and the other proteins is the absence of an arginine residue. Tyrosine 39 of Ran is in a similar location to the *cis*-arginine of G α and the *trans*-arginines from RasGAP and the RhoGAPs. The maximal rates for GAP-stimulated GTPase reactions are very similar, with a k_{cat} of 5–30 s⁻¹ irrespective of the presence or absence of an arginine, which raises the issue of the reaction mechanism.

The function of GAPs, on the basis of evidence for Ras and Rho, is to supply catalytic residues into the active site and stabilize the catalytic machinery. A catalytic glutamine is conserved throughout all G α and Ras-like proteins except Rap, and superimposes well in the available structures (Fig. 4a). This residue is important for catalysis because the conventionally used glutamine to leucine

mutation, Gln61Leu in Ras¹⁸ and Gln69Leu in Ran¹⁹, has a lower intrinsic GTPase reaction that is not stimulated by GAPs. To highlight the importance of glutamine for the reaction and to exclude the possibility that the glutamine to leucine mutation interferes sterically with catalysis, we investigated the properties of a Gln69Ala mutation in Ran. This mutant also has a lower intrinsic GTPase activity and is stimulated only 1.5-fold at saturating RanGAP concentrations, whereas wild-type Ran is stimulated 10⁵-fold (Supplementary Information). This 10⁶-fold increase in overall GTPase rate, in which is also observed for Ras, highlights the crucial importance of glutamine.

In the structures of GNBPs-GAP complexes mentioned above, glutamine was shown to form a double hydrogen bond with the nucleophilic water molecule and one of the oxygens of the γ -phosphate. Its position was stabilized by a hydrogen bond to the main-chain CO of the catalytic arginine. In the ternary complex, Gln 69 of Ran forms a hydrogen bond with Tyr 39 of Ran and with Asn 131 of GAP, and is also constrained by van der Waals interactions with the loop that contains Gly 133 of RanGAP (Fig. 3a). It overlays well with the catalytic glutamine of the other structures. In the Ran-importin- β complex, Gln 69 of Ran is fixed such that it cannot point towards the catalytic water, in accordance with the observation that GTP hydrolysis is inhibited in these complexes (Fig. 4b). The glutamine residue is again kept in a catalysis-incompetent state in the binary Ran-RanBD complex, in line with biochemical evidence. The role of RanGAP in promoting GTP hydrolysis on Ran is thus to stabilize the switch II region and orient its catalytic glutamine. This situation is similar to the G α proteins in which the stabilization of loops is the most important effect of GTPase stimulation¹⁷ by RGS.

In contrast to glutamine, the catalytic arginine is not equally important for GTPases. Its mutation completely abolishes activity in RabGAP²⁰, and reduces catalysis 2,000-fold in rasGAP²¹, 50–240-fold in RhoGAP^{16,22}, and at least 40-fold in G α proteins²³. In the case of Ran, the structural and biochemical evidence supports the idea that arginine is not involved. Assuming that the most important element of catalysis is the presence of a correctly placed and stabilized glutamine residue, we can speculate about the difference that the presence or absence of an arginine makes to the phosphoryl transfer of GTP-binding proteins. It has been proposed that a positively charged residue that contacts the transferred phosphate in the transition state is anticatalytic for a dissociative mechanism^{1,24}; therefore, we might argue that whereas the structural evidence for Ras-RasGAP and Rho-RhoGAP points to a more associative nature of the transition state with a pentacoordinated γ -phosphate, Ran-RanGAP might favour a more dissociative transition state with a metaphosphate-like configuration. Whether or not the strongly negative electrostatic surface of RanGAP contributes to the overall rate enhancement and favours the proposed mechanism remains to be established. Fourier transform infrared spectroscopy, which has been used successfully to monitor the neurofibromin-stimulated GTPase reaction of Ras²⁵, could be used to follow the charge distribution on the phosphates during the reaction.

In summary, we have shown that RanBP has only an indirect role in RanGAP-mediated GTP hydrolysis on Ran. Contrary to expectations, the mechanism of GTP hydrolysis does not involve an arginine finger and RanGAP does not directly contact the active site. Instead, we propose that the correct positioning of the catalytic machinery on Ran itself and its shielding from the solvent are sufficient for the phosphoryl transfer reaction. As we have found no essential arginine in RapGAP by mutational analysis (A.W., unpublished data) and the guanylate-binding protein GBP1 has no arginine or glutamine close to the active site²⁶, we should be prepared for many different GAP mechanisms in GTP-binding proteins. □

Methods

Protein purification and crystallization

We prepared RanGAP from *Schizosaccharomyces pombe* and human Ran as described. Human RanBP1 was expressed in BL21 *Escherichia coli* as a glutathione *S*-transferase fusion protein and subsequently purified with Glutathione–Sepharose 4 Fast Flow (Amersham Pharmacia). We loaded RanBP1 onto an affinity column and added Ran–GppNHp in twofold excess afterwards. The binary complex was eluted after a 12-h incubation with thrombin protease. We then added RanGAP in excess and purified the ternary complex by size-exclusion chromatography in buffer (20 mM Tris–HCl pH 7.5, 2 mM MgCl₂, 2 mM dithioerythritol (DTE)). Crystals were grown by hanging drop from 18.5–20.5% PEG4000, 100 mM potassium acetate, 100 mM Tris–HCl pH 7.5 for the GppNHp complex. After 3–4 d at 20 °C, thin needle protein crystals appeared (500 × 45 × 45 μm³).

For the transition-state analogue complex, RanBP1, Ran–GDP and RanGAP were mixed in a buffer containing 2 mM Al³⁺ and 30 mM NaF. Subsequently, the ternary complex was purified by size-exclusion chromatography in 20 mM Tris–HCl pH 7.5, 2 mM MgCl₂, 2 mM DTE containing additional 30 mM NaF. Crystals were grown from 18.5–20.5% PEG4000, 100 mM Tris–HCl pH 7.5 at 20 °C.

Data collection and structure determination

The crystals, with space group *P*1 and unit-cell dimensions of *a* = 101.6 Å, *b* = 103.1 Å, *c* = 120.2 Å, α = 71.6, β = 80.6, γ = 67.8, were very sensitive to radiation damage. Data sets were obtained by translating a single crystal needle and successively exposing small sections to the beam. Data were collected on MAR image plate detectors at the beam lines ID13 and ID14-1 (for AlF₃) at the European Synchrotron Radiation Facility (ESRF) and were processed with DENZO²⁷. We solved the structure of the complex by molecular replacement using AMORE²⁸ with RanGAP⁷ as a model (correlation coefficient 0.41, *R*-factor 0.47). The initial electron density was improved greatly by four-fold averaging using the non-crystallographic symmetry (NCS) in the unit cell. After inclusion of Ran and RanGAP the DM²⁸ correlation coefficient for the RanBP1 electron density was 0.77. The model of the ternary complex was built using the program O²⁹ and refined with CNS applying NCS restraints³⁰. The final model comprises residues 2–345 of RanGAP, 8–213 of Ran and 22–167 of RanBP1.

Received 9 August; accepted 29 November 2001.

- Scheffzek, K. *et al.* The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. *Science* **277**, 333–338 (1997).
- Rittinger, K. *et al.* Structure at 1.65 Å of RhoA and its GTPase-activating protein in complex with a transition-state analogue. *Nature* **389**, 758–762 (1997).
- Görlich, D. & Kutay, U. Transport between the cell nucleus and the cytoplasm. *Annu. Rev. Cell Dev. Biol.* **15**, 607–660 (1999).
- Bischoff, F. R., Krebber, H., Smirnova, E., Dong, W. & Ponstingl, H. Co-activation of RanGTPase and inhibition of GTP dissociation by Ran-GTP binding protein RanBP1. *EMBO J.* **14**, 705–715 (1995).
- Bischoff, F. R. & Görlich, D. RanBP1 is crucial for the release of RanGTP from importin beta-related nuclear transport factors. *FEBS Lett.* **419**, 249–254 (1997).
- Becker, J. *et al.* RNA1 encodes a GTPase-activating protein specific for Gsp1p, the Ran/TC4 homologue of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **270**, 11860–11865 (1995).
- Hillig, R. C. *et al.* The crystal structure of rna1p: a new fold for a GTPase-activating protein. *Mol. Cell.* **3**, 781–791 (1999).
- Vetter, I. R., Arndt, A., Kutay, U., Görlich, D. & Wittinghofer, A. Structural view of the Ran-Importin β interaction at 2.3 Å resolution. *Cell* **97**, 635–646 (1999).
- Chook, Y. M. & Blobel, G. Structure of the nuclear transport complex karyopherin- β 2-Ran.GppNHp. *Nature* **399**, 230–237 (1999).
- Haberland, J., Becker, J. & Gerke, V. The acidic C-terminal domain of rna1p is required for the binding of Ran.GTP and for RanGAP activity. *J. Biol. Chem.* **272**, 24717–24726 (1997).
- Lounsbury, K. M., Richards, S. A., Carey, K. L. & Macara, I. G. Mutations within the Ran/TC4 GTPase. Effects on regulatory factor interactions and subcellular localization. *J. Biol. Chem.* **271**, 32834–32841 (1996).
- Vetter, I. R., Nowak, C., Nishimoto, T., Kuhlmann, J. & Wittinghofer, A. Structure of a Ran-binding domain complexed with Ran bound to a GTP analogue: implications for nuclear transport. *Nature* **398**, 39–46 (1999).
- Richards, S. A., Lounsbury, K. M. & Macara, I. G. The C terminus of the nuclear RAN/TC4 GTPase stabilizes the GDP-bound state and mediates interactions with RCC1, RAN-GAP, and HTF9A/RANBP1. *J. Biol. Chem.* **270**, 14405–14411 (1995).
- Scheffzek, K., Klebe, C., Fritz-Wolf, K., Kabsch, W. & Wittinghofer, A. Crystal structure of the nuclear Ras-related protein Ran in its GDP-bound form. *Nature* **374**, 378–381 (1995).
- Haberland, J. & Gerke, V. Conserved charged residues in the leucine-rich repeat domain of the Ran GTPase activating protein are required for Ran binding and GTPase activation. *Biochem. J.* **343**, 653–662 (1999).
- Nassar, N., Hoffman, G. R., Manor, D., Clardy, J. C. & Cerione, R. A. Structures of Cdc42 bound to the active and catalytically compromised forms of Cdc42GAP. *Nature Struct. Biol.* **5**, 1047–1052 (1998).
- Tesmer, J. J., Berman, D. M., Gilman, A. G. & Sprang, S. R. Structure of RGS4 bound to AlF₄[–]-activated G_{12i}: stabilization of the transition state for GTP hydrolysis. *Cell* **89**, 251–261 (1997).
- Der, C. J., Finkel, T. & Cooper, G. M. Biological and biochemical properties of human rasH genes mutated at codon 61. *Cell* **44**, 167–176 (1986).

- Klebe, C., Bischoff, F. R., Ponstingl, H. & Wittinghofer, A. Interaction of the nuclear GTP-binding protein Ran with its regulatory proteins RCC1 and RanGAP1. *Biochemistry* **34**, 639–647 (1995).
- Albert, S., Will, E. & Gallwitz, D. Identification of the catalytic domains and their functionally critical arginine residues of two yeast GTPase-activating proteins specific for Ypt/Rab transport GTPases. *EMBO J.* **18**, 5216–5225 (1999).
- Ahmadian, M. R., Stege, P., Scheffzek, K. & Wittinghofer, A. Confirmation of the arginine-finger hypothesis for the GAP-stimulated GTP-hydrolysis reaction of Ras. *Nature Struct. Biol.* **4**, 686–689 (1997).
- Graham, D. L., Eccleston, J. F. & Lowe, P. N. The conserved arginine in rho-GTPase-activating protein is essential for efficient catalysis but not for complex formation with Rho.GDP and aluminium fluoride. *Biochemistry* **38**, 985–991 (1999).
- Berman, D. M., Wilkie, T. M. & Gilman, A. G. GAIP and RGS4 are GTPase-activating proteins for the G_i subfamily of G protein α subunits. *Cell* **86**, 445–452 (1996).
- Maeagley, K. A., Admiraal, S. J. & Herschlag, D. Ras-catalyzed hydrolysis of GTP: a new perspective from model studies. *Proc. Natl Acad. Sci. USA* **93**, 8160–8166 (1996).
- Allin, C., Ahmadian, M. R., Wittinghofer, A. & Gerwert, K. Monitoring the GAP catalyzed H-Ras GTPase reaction at atomic resolution in real time. *Proc. Natl Acad. Sci. USA* **98**, 7754–7759 (2001).
- Prakash, B., Renault, L., Praefcke, G. J., Herrmann, C. & Wittinghofer, A. Triphosphate structure of guanylate-binding protein 1 and implications for nucleotide binding and GTPase mechanism. *EMBO J.* **19**, 4555–4564 (2000).
- Otwinski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- Collaborative Computational Project No. 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Jones, T. A. & Kjeldgaard, M. Electron-density map interpretation. *Methods Enzymol.* **277**, 173–208 (1997).
- Brunger, A. T. *et al.* Crystallography and NMR system (CNS): A new software system for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank M. Alt for atomic absorption spectroscopy; staff of the European Molecular Biology Laboratory/ESRF for access and support at beam lines ID13 and ID14; R. Hillig for the coordinates of uncomplexed RanGAP before release; and the Deutsche Forschungsgemeinschaft (DFG) for a grant to I.R.V.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.W. (e-mail: alfred.wittinghofer@mpi-dortmund.mpg.de). Coordinates have been deposited with the Protein Data Bank (accession numbers 1K5D and 1K5G).

erratum

Formation of coastline features by large-scale instabilities induced by high-angle waves

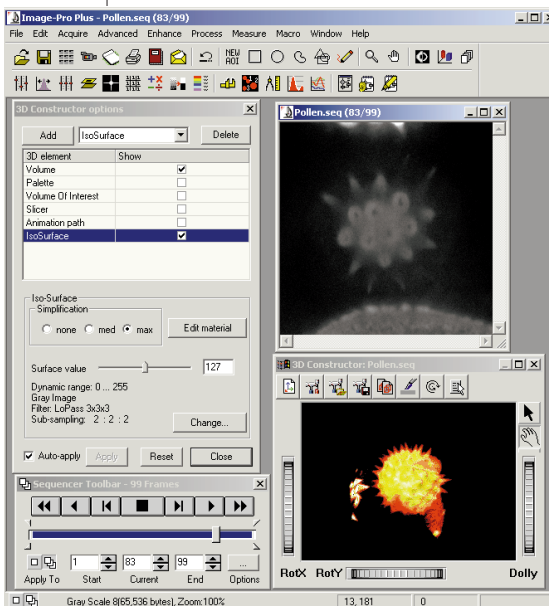
Andrew Ashton, A. Brad Murray & Olivier Arnault

Nature **414**, 296–300 (2001).

In this Letter, the last name of the third author, Olivier Arnault, was misspelled as 'Arnault'. □

Sharpen up your image

Featuring image analysis, capture and sharing.



Grain of truth: pollen in Image-Pro Plus.

Image-Pro Plus plug-ins

MediaCybernetics www.mediacy.com

Increase your exposure

The 3D Constructor, SharpStack and AFA plug-ins from MediaCybernetics can be used with Image-Pro Plus version 4.5 to achieve six-dimensional acquisition, enhancement, management and analysis. 3D Constructor can be used to render Z-stacks to expose relationships between and within objects. It allows users to rotate in X, Y, and Z dimensions, zoom, change background colour, view numerous isosurface rendering options, animate by selecting rotation angles, select volume of interest and display orthogonal and oblique slices. SharpStack extracts clear, sharp images from a stack of hazy planes using nearest neighbour, no neighbour and inverse algorithm functions to sharpen one or all planes from a Z-stack. AFA allows Image-Pro users to manage all combinations of acquisition modes: time, channel, focus and stage position. It includes a set manager that enables users to organize, view and extract any single mode sequence from a set and allows for X/Y tiling to achieve seven-dimensional analysis.

iVISION

BioGenex www.biogenex.com

Automated digital image analysis

The BioGenex iVISION is a colour, high-resolution, high-sensitivity, automated, and digital image analysis system. The iVISION comes with proprietary digital imaging and automa-

tion technologies offering, say BioGenex, more than four times the resolution and about two orders of magnitude higher detection sensitivity than other image analysis systems in the market. In addition, the iVISION system is equipped with auto slide-loaders that handle random access loading and unloading capabilities for up to 50 bar-coded slides. The iVISION system can be used for high-speed scanning and detection, acquisition and archiving, rare cell detection, quantitation, data/image retrieval, tissue microarray applications. The iVISION system can be used for digital archival, quantitation for therapy selection, multi-parameter target detection, and consultation.

ZetaPALS

Brookhaven Instruments Corporation

www.bic.com

Potential value

The ZetaPALS instrument can be used to determine the size and zeta potential of particles in high ionic strength solutions. It uses phase analysis light scattering (PALS), a technique that is said to be 1,000 times more sensitive than conventional methods that depend on measuring shifts in light frequency, to produce determinations of zeta potential in solutions that are highly conductive.

PARISS spectrometer system

Lightform www.lightforminc.com

Plain sailing

LightForm's PARISS (prism and reflector imaging spectrometer system) features wavelength dispersive spectral imaging. Rather than acquiring wavelengths one at a time through filters, this system acquires all wavelengths simultaneously and then applies algorithms to analyse the data. It automatically associates spectral characteristics with objects, labels, tags, molecules or physiological conditions. Slight changes in transmission, absorption, reflection or fluorescence characteristics that occur as a result of subtle alterations in physiological conditions, pH, defects, FRET and FRAP are detected. Objects or areas occupying less than 1 μm in diameter can be characterized. The system generates a topographical map of areas that meet or deviate from

expected criteria. It can be used with a variety of heterogeneous samples, including cells or any material that presents a complex, spatially resolved mixture of constituents.

UltraVIEW

Perkin Elmer www.perkinelmer.com

Fast confocal imager for live cell analysis

This new version of the UltraVIEW live cell imager incorporates a faster camera with higher sensitivity and comes with Improvion's Velocity three-dimensional imaging package for multicolour volume rendering. UltraVIEW captures real-time images of the complex pathways of cell biology without causing photo damage to the proteins and cellular components. It can be used for *in vitro* analysis of proteins identified as disease targets.

DN100

Nikon www.ave.nikon.co.jp

Real-time image sharing

Nikon's DN100 digital camera's networking feature allows microscope images to be shared over a local area network in real time, facilitating the movement of microscope slides between laboratories and hospitals or factories and laboratories. The system con-



High browse: use Nikon's DN100 via Netscape.

sists of a camera and control unit with cable to a monitor screen. A PC card slot is provided so that image data can be uploaded, stored and transported. Networked parties can view images at any time, including during capture, using a standard browser. Images can be captured at up to 15 frames per second and parameters such as exposure, white balance, colour, noise reduction and shading compensation can be adjusted by anyone on the network. Other features include an electronic zoom (from $1.4\times$ to $16\times$ in eight steps), storage options, line drawing and viewing of images side by side.

VersaDoc

Bio-Rad www.discover.bio-rad.com

Smooth operator

The VersaDoc quantitative imaging system uses flat fielding technology for both ultraviolet and white light illumination to eliminate non-uniformities caused by variations in the optical path. It has a CV < 5% and can be used to capture high-resolution digital images from single and multicolour fluorescence, chemiluminescence, chemifluorescence and calorimetric samples. VersaDoc is suitable for use with electrophoretic and microplate samples, among others.



What's up VersaDoc?

Camedia E-10 camera

Olympus www.olympus-europa.com

Sharp and versatile

The Camedia E-10 features a 4.0 megapixel CCD with a $4\times$ F2.0–2.4 optical zoom lens. It can be fitted to a microscope with a C-mount adaptor or used alone with larger specimens or in field studies. The camera has three light metering systems — spot metering, recommended for fluorescence samples, centre-weighted average and ESP. Exposure can be chosen manually or with a choice of aperture priority, shutter priority or completely automated. Exposure compensation can be finely adjusted and auto-bracketing allows three sequential frames to be taken about the mean exposure. Repeated exposure at up to three frames per second or interval shooting from one minute to 24 hours is a useful feature for studies of growth or cell change. Resolution can be set up to 2240×1680 pixels and down to 640×480 thumbnails with a choice of setting in between. Images can be stored on up to 64 MB SmartMedia cards or downloaded directly to a PC, bypassing the camera's image processing system.

DC 500

Leica www.leica-microsystems.com

It's all in the detail

The DC 500 digital camera from Leica can acquire images in with resolutions greater than 12 megapixels at 12 bits per colour, producing highly detailed images. An integrated Peltier cooling system and heat dissipating metallic casing enhance the camera's dynamic range and reduce signal-to-noise. A six-minute exposure time improves image cap-

ture in low-light applications. Incorporation of Firewire technology allows transfer of images to a computer at up to 400 megabits per second and saved in multiple graphical formats. Leica's IM 50 image management software is included, allowing users to archive and preview images in an image gallery.

GeneTools Version 3.1

Syngene www.syngene.com

Gel analysis, one row at a time

This new version of GeneTools includes an automated multi-layer gel analysis capability for screening large numbers of DNA samples. It is particularly suitable for clinical or plant geneticists screening different DNA samples and markers with several rows of wells in one gel. The software will automatically analyse each row as if it were a separate gel, thus obtaining results from several small gels using only one large gel and a single electrophoresis run. GeneTools Version 3.1 comes as standard with Syngene's ChemiGenius and GeneGenius. Applications include one-dimensional and spot blot analysis, and colony counting.

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENTS



**Nail Polish is for Fingernails.
Rubber Cement is for Paper.**



**Hybridization Chambers
are for Slides.**

For the Best Results, Use the Right Tool.

- Provide uniform staining and binding
- Prevent sample drying and damage
- For microarrays, immunocytochemistry, *in situ* hybridization, confocal imaging, and histochemistry

Call us today at 1-800-245-4024 or visit www.arraying.com for more details.

S&S – Easy to Work With
GRACE BIO • LABS
Schleicher & Schuell
BioScience

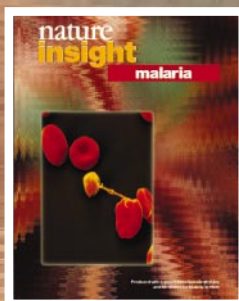
USA • Tel (603) 352-3810 • techserv@s-and-s.com
Germany • Tel +49-5561-791-575 • lifescience@s-and-s.de



Exposed: Camedia E-10 *in situ*.

nature insight

malaria



Cover illustration
Coloured scanning electron micrograph of a *Plasmodium falciparum* protozoan with red blood cells (courtesy of Eye of Science/SPL).

In any given year, nearly ten per cent of the global population will suffer from malaria — 500 million clinical cases — and more than 1 million will die as a result; a death from malaria every 30 seconds. In Africa, the disease kills one child in twenty before five years of age.

And things are getting worse, as malaria is undergoing a resurgence. The main contributing factors are the emergence of drug-resistant strains of the parasite, the appearance of mosquitoes (which transmit the parasite) that are resistant to insecticides, environmental changes and increased population.

After decades of relative lack of attention, growing international awareness and funding has led to new efforts towards controlling the disease. A wide-ranging coalition of interests is being marshalled to combat malaria, including: global healthcare planning, specifically the World Health Organization's 'Roll Back Malaria' campaign, which aims to halve the burden of disease by 2010; pharmaceutical industry support; research coordination, with the main funding agencies coming together in the Multilateral Initiative on Malaria; and philanthropy, most notably the Malaria Vaccine Initiative supported by the Bill and Melinda Gates Foundation.

While this points to an appropriate level of concern and urgency, the most critical challenge is to the researchers. Can they supply the knowledge and tools needed to combat this devastating disease? This Insight will go a long way to answering that question, describing latest research developments, likely future progress and the practical impact that the new knowledge will have.

We are pleased to acknowledge the financial support of GlaxoSmithKline and Medicines for Malaria Venture in producing this Insight. As always, *Nature* carries sole responsibility for all editorial content and peer review.

Ursula Weiss Senior Editor

introduction:

- 670 Malaria in 2002**
*B. Greenwood
& T. Mutabingwa*

review articles:

- 673 The pathogenic basis of malaria**
*L. H. Miller, D. I. Baruch,
K. Marsh & O. K. Doumbo*
- 680 The economic and social burden of malaria**
J. Sachs & P. Malaney
- 686 Medical need, scientific opportunity and the drive for antimalarial drugs**
R. G. Ridley
- 694 Progress and challenges for malaria vaccines**
T. L. Richie & A. Saul
- 702 *Plasmodium*, human and *Anopheles* genomics and malaria**
*S. L. Hoffman,
G. M. Subramanian,
F. H. Collins & J. C. Venter*
- 710 Satellite imagery in the study and forecast of malaria**
*D. J. Rogers,
S. E. Randolph,
R. W. Snow & S. I. Hay*

Editor, *Nature*: **Philip Campbell**
Insights Editor: **Lesley Anson**
Editorial Assistant: **Simon Gibson**
Production Editor: **Simon Gribbin**

Art Director: **Majo Xeridat**
Diagrams: **Ann Thomson**
Vicky Askew
Suzanne Coleman

Production Manager: **Yvonne Strong**
Sponsorship: **Emma Jones**

Malaria in 2002

Brian Greenwood* & Theonest Mutabingwa*†

*Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

†National Institute for Medical Research, PO Box 9653, Dar es Salaam, Tanzania

The burden of malaria is increasing, especially in sub-Saharan Africa, because of drug and insecticide resistance and social and environmental changes. Thus, there is an urgent need for vaccines, new drugs and insecticides. Parasite, mosquito and human genome projects are helping in the search for new control tools and international donors are developing new funding mechanisms that could make them available to poor countries. But these new tools will achieve their maximum impact only if additional resources are deployed to strengthen malaria research and control communities in countries where the new tools will be used.

Malaria was eliminated from the United States and from most of Europe during the first half of the twentieth century as a result of changes in land use, agricultural practices and house construction and some targeted vector control. The development of the highly effective, residual insecticide DDT initiated a global eradication programme in the 1950s and 1960s which was initially very successful in many countries such as India, Sri Lanka and the former Soviet Union. However, this success was not sustained because of the costs of the programme, the resistance of many communities to repeated spraying of their houses and the emergence of resistance to DDT. Furthermore, with the exception of a few pilot schemes, no sustained effort was made to control malaria in sub-Saharan Africa.

The elimination of malaria from most of Europe and from North America and the failure of the global malaria eradication programme led to a loss of interest in malaria for a period of about 25 years from the early 1970s to the late 1990s. Only 3 of 1,223 new drugs developed during the period 1975–1996 were antimalarials¹. Industry lost interest in the development of insecticides for public health use and support for research on malaria declined. Furthermore, in many malaria-endemic countries, national malaria control programmes, established during the colonial period and sustained during the period when elimination of malaria was considered to be an achievable goal, collapsed. Thus, for many years, there was little change in mortality and morbidity from malaria, especially in Africa. Recently, the malaria situation has deteriorated and mortality from malaria is probably increasing in sub-Saharan Africa. Some of the reasons for this are shown in Box 1. By far the most important is the development by *Plasmodium falciparum* of resistance to cheap and effective drugs such as chloroquine and sulphadoxine/pyrimethamine.

The burden of malaria

Although malaria remains an important health problem in some parts of Asia and South America, its main impact is in sub-Saharan Africa where at least 90% of deaths from malaria occur. The most important reason for the persistence of malaria in Africa is the presence of the vector *Anopheles gambiae*, although social and economic factors are also important. *A. gambiae* feeds preferentially on humans and is long-lived, making it particularly effective at transmitting malaria from one person to another. The entomological inoculation rate (EIR), a measure of the frequency with which an individual is bitten by an infectious mosquito,

rarely exceeds 5 per year in Asia or South America. In contrast, EIRs of over 1,000 have been recorded in several parts of sub-Saharan Africa. In savannah areas of West Africa, it is not unusual to collect in one room during the course of a single night several hundred mosquitoes of the *A. gambiae* complex, 1–5% of which are infectious. The task of interrupting transmission in such a situation is daunting.

There are major differences in the prevalence of malaria between countries in Africa, between districts in the same country and even between villages situated only a mile or two apart (Fig. 1). Consideration of the whole of tropical Africa as an area of hyper-endemic malaria transmission is a simplification of a very complex epidemiological situation. Fortunately, new techniques, including satellite imagery, are making it possible to construct much more accurate maps of the distribution of malaria in Africa than have been available in the past (see review in this issue by Rogers *et al.*, pages 710–715). Superimposition of plots of population density and of the distribution of health facilities on maps of malaria risk may identify imbalances between needs and resources. For example, a study in Kenya showed that insecticide-treated bednet (ITN) programmes were concentrated in areas where nongovernmental organizations had strong links, rather than in areas where malaria risk was greatest².

Assessing the burden of malaria accurately is difficult because most deaths from malaria occur at home, the clinical features of malaria are very similar to those of many other infectious diseases and good quality microscopy is available in only a few centres. Despite these limitations, attempts have been made recently to estimate the global burden of malaria more accurately than had been done previously^{3,4}. It is now believed that at least one million deaths occur from malaria each year — a death from malaria every 30 seconds. However, when successful interventions have been introduced into malaria-endemic areas, their impact on overall mortality has usually been more marked than would have been expected. Thus, an effective intervention, such as a highly protective malaria vaccine, might save many more deaths than expected.

In many parts of Africa, young children experience several clinical attacks of malaria each year. Fortunately, only a small proportion of these attacks (1–2%) leads to severe complications. Why some children infected with *P. falciparum* develop a life-threatening illness whereas others experience only a febrile illness is still not fully understood. Parasite, host genetic and socioeconomic factors are all likely to have a role (see review in this issue by Miller *et al.*, pages 673–679). Deaths from severe malaria anaemia are increasing because of drug resistance and concerns over

Box 1

Factors contributing to the increasing burden of malaria

Drug resistance. In parts of Southeast Asia, *P. falciparum* is now resistant to almost all antimalarial drugs and strains of chloroquine-resistant *P. vivax* have emerged⁶. In Africa, chloroquine resistance is widespread and resistance to sulphadoxine/pyrimethamine is being detected increasingly frequently (ref. 7; and see review in this issue by Ridley, pages 686–693). Spread of highly resistant parasites of the type found in Southeast Asia to Africa could herald a medical disaster⁸.

Insecticide resistance. The emergence of mosquitoes resistant to pyrethroid insecticides in West and South Africa now threatens insecticide-treated bednet programmes⁹.

War and civil disturbance. Wars in Africa and elsewhere have led to major problems of malaria transmission in refugee populations. Collapse of health services has led to a resurgence of malaria in some parts of the former Soviet Union¹⁰.

Environmental changes. In Thailand, malaria vectors have established themselves in rubber plantations¹¹ and construction of small dams has led to an increase in malaria transmission in Ethiopia¹².

Climatic changes. Global warming may have contributed to the spread of malaria into some previously malaria-free mountainous areas of Asia, Africa and South America, but other environmental factors are likely to be involved in this process¹³. Floods associated with increased El Niño rains have precipitated epidemics of malaria in Africa¹⁴.

Travel. About 7,000 imported cases of malaria are recorded in Europe each year¹⁵. In malaria-endemic countries, travel from malaria-free areas to malaria-endemic areas for work is probably an increasingly important, and largely unrecognized, cause of severe malaria¹⁶.

Population increase. During the past two decades the population of many malaria-endemic countries has doubled, thus greatly increasing absolute numbers of those at risk.

transfusion in communities where the prevalence of infection with human immunodeficiency virus (HIV) is high. Mortality from cerebral malaria remains at about 20%, even in hospitals with good facilities, and about 10% of children who survive are left with neurological defects which may interfere with their education and subsequent chances of employment⁵.

Malaria is especially dangerous during pregnancy when some of the immunity acquired by adults as a result of repeated infections is lost. In areas of low transmission, malaria in pregnancy may lead to death of the mother, abortion of the fetus or a stillbirth. In areas of high malaria transmission, heavy parasitization of the placenta with *P. falciparum* may occur (see review in this issue by Miller *et al.*), especially in early pregnancies, impairing placental function and resulting in a low birth weight baby. Because infant survival is so sensitive to even small reductions in birth weight, malaria infection during pregnancy almost certainly increases infant mortality, although this has never been demonstrated directly.

The direct economic costs of malaria that result from treatment and from time away from work or school are enormous, but the overall economic impact of malaria is likely to be much more substantial than suggested by estimates of direct costs alone (see review in this issue by Sachs and Malaney, pages 680–685).

A more hopeful future?

The past five years has seen a pronounced re-awakening of interest in malaria in the richer countries of the world. Statements on the need for greater efforts to control malaria have been made at a number of high-profile political meetings in Africa and in industrialized countries.



Figure 1 The risk of malaria varies from area to area across Africa. In areas of relatively low endemicity, such as highland areas of Tanzania, the whole family is at risk. (Figure courtesy of C. Drakeley.)

Research scientists have had an important catalytic role in this process. In 1997, a meeting was held in Dakar, Senegal, attended by most of the small number of scientists undertaking malaria research in Africa and by their sponsors, which established priorities for a multidisciplinary programme of research. This meeting was held under the auspices of a new organization — the Multilateral Initiative on Malaria (MIM) — which has subsequently supported a number of important initiatives in research, training and information technology.

Since the Dakar meeting, increased funds for malaria research have been provided by established donors such as the National Institutes for Health and the Wellcome Trust, and by new donors such as The Gates Foundation. Another important new initiative has been the creation by the World Health Organization of the Roll Back Malaria (RBM) initiative, a partnership that is drawing together all the main groups interested in malaria control at international and national levels. RBM has raised the profile of malaria control during the past two years and it has set itself ambitious targets. These include achieving 60% access to effective treatment, protection during pregnancy, use of ITN or other appropriate measures of vector control by at-risk groups across Africa by 2005, and a reduction in global mortality from malaria by a half by 2010. RBM now faces the difficult task of delivering on the expectations that have been raised.

The population at risk from malaria in Africa is large, around 500 million, and increasing rapidly; thus the costs of even a basic control programme that will cover the whole population will be substantial, perhaps as much as US\$2 billion each year for an indefinite period. Where might this come from? A part might come from the users, as even poor people are prepared to pay for an intervention against mosquitoes (and malaria) that they perceive to be effective, such as a bednet, and substantial additional sums are spent on insecticide coils and other repellents. However, the poor who are at greatest risk are likely to remain unprotected. Additional funds could come from the government through the redistribution of budgets from unproductive areas and from funds released by debt relief. But there is little doubt that malaria will not be controlled effectively in the poorest countries of Africa unless financial help is provided by wealthy, industrialized countries. The recent creation of a Global Fund to support the purchase of goods needed for the management of HIV, tuberculosis and malaria is a promising step forward but it is unlikely to be sufficient by itself, especially if there are no parallel improvements in the healthcare systems that will be needed to deliver these commodities in the most effective way.

Are there adequate tools for the control of malaria in highly endemic areas if substantial additional funds were to be made available? There is little doubt that much could be achieved using the

Box 2

New developments in methods for malaria control

Vector control. The insecticide DDT, still a valuable control tool in some malaria situations when used for household spraying, has been saved from a global ban¹⁷.

Interruption of human-vector contact. Insecticide-treated bednets and curtains have been shown to provide significant protection against malaria in almost all epidemiological situations¹⁸. Methods are being developed for incorporating insecticide into netting materials, so as to obviate the need for repeated re-treatment, and for increasing bednet usage¹⁹.

Intermittent treatment for malaria in pregnancy. Administration of sulphadoxine/pyrimethamine once during the second and once during the third trimester of pregnancy protects pregnant women against severe malaria anaemia and improves birth weight, although HIV-positive women may require more frequent treatment^{20–22}.

Intermittent treatment in infancy. A recent study in Tanzania showed that treatment of infants with Fansidar at the time of vaccination reduced the incidence of malaria and anaemia substantially²³.

Improving access to treatment. As a result of training mothers how to treat malaria correctly, mortality was reduced substantially in Ethiopian children²⁴.

Improved compliance with treatment. It has been shown that training shopkeepers on the importance of selling a full course of treatment and packaging of tablets can help to ensure that full courses of treatment are given^{25,26}.

Development of artemisinin suppositories. Drugs derived from the plant *Artemisia annua* are effective when given as a suppository, thus providing a means of initial treatment for cases of severe malaria in peripheral clinics where injections cannot be given.

Combination therapy. There is increasing recognition that, as far as possible, antimalarials should be used in combination so as to improve therapeutic efficacy and to decrease the chance of resistant parasites emerging. However, the benefits of combination therapy must be balanced against the draw-back of increased costs²⁷.

existing tools, which have been shown to be at least partially effective (Box 2). The success of Vietnam in reducing mortality and morbidity from malaria by using ITN and artemisinin-based treatment of clinical cases serves as a model for what can be done.

The *Plasmodium* genome project (see review in this issue by Hoffman *et al.*, pages 702–709) is identifying potential new drug targets at an accelerating rate and this knowledge may allow a drug already in use for some other purpose to be developed as an antimalarial quickly and relatively cheaply (see review in this issue by Ridley, pages 686–693). In addition, a variety of public-private partnerships are being set up to develop drugs for poor countries in a more effective way than has been the case in the past, as exemplified by the development of the new antimalarial Lapdap (chlorproguanil/dapsone). Exploitation of the *Plasmodium* genome project will facilitate vaccine development (see review in this issue by Richie and Saul, pages 694–701) and the anopheline genome project may help to identify novel insecticides, ways of overcoming insecticide resistance and facilitate attempts to develop mosquitoes that are resistant to malaria infection (see review in this issue by Hoffman *et al.*). Thus, at a technical level, prospects for the control of malaria have never been brighter.

Capacity development in endemic countries

Effective national malaria control programmes must be able to detect outbreaks, to monitor parasite and mosquito populations for

changes in drug and insecticide resistance, and to make changes in treatment and control guidelines in a rapid and responsible way based on sound scientific findings. Thus, operation of an effective malaria control programme requires trained staff at many levels, ranging from traditional healers and shopkeepers to programme managers. In addition, it is essential that developing country scientists should be given the training and resources needed to allow them to participate fully in the exciting developments that are taking place in many areas of malaria research. Unfortunately, few malaria-endemic countries have given sufficient support to their research scientists or malaria control teams in recent years. So far relatively little attention has been paid by the principal donors who are providing increasing levels of support for malaria research and control to capacity development in malaria-endemic countries. A major investment in staff recruitment, training and support after training is needed if the new tools that are being developed for malaria control are to be introduced and evaluated effectively in the areas where they are needed most. □

1. Trouiller, P. & Olliaro, P. L. Drug development output from 1975 to 1996: what proportion for tropical diseases? *Int. J. Infect. Dis.* **3**, 61–63 (1998).
2. Shretta, R., Omumbo, J. & Snow, R. W. Community based healthcare and its relationship to insecticide treated bednets in Kenya. Report for the Kenyan Ministry of Health (1998).
3. Snow, R. W., Craig, M., Deichmann, U. & Marsh, K. Estimating mortality, morbidity and disability due to malaria among Africa's non-pregnant population. *Bull. World Health Org.* **77**, 624–640 (1999).
4. Breman, J. G. The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. *Am. J. Trop. Med. Hyg.* **64**(Suppl.), 1–11 (2001).
5. Holding, P. A. & Snow, R. W. Impact of *Plasmodium falciparum* malaria on performance and learning: review of the evidence. *Am. J. Trop. Med. Hyg.* **64**(Suppl.), 68–75 (2001).
6. Fryauff, D. J. *et al.* Chloroquine-resistant *Plasmodium vivax* in transmigration settlements of West Kalimantan, Indonesia. *Am. J. Trop. Med. Hyg.* **59**, 513–518 (1998).
7. Mutabingwa, T. *et al.* Chlorproguanil-dapsone for treatment of drug-resistant falciparum in Tanzania. *Lancet* **358**, 1218–1223 (2001).
8. White, N. J. *et al.* Averting a malaria disaster. *Lancet* **353**, 1965–1967 (1999).
9. Chandre, F. *et al.* Status of pyrethroid resistance in *Anopheles gambiae* sensu lato. *Bull. World Health Org.* **77**, 230–234 (1999).
10. Pitt, S. *et al.* War in Tajikistan and re-emergence of *Plasmodium falciparum*. *Lancet* **352**, 1279 (1998).
11. Gomes, M., Linthicum, K. & Haile, M. in *New and Resurgent Infections* (eds Greenwood, B. & De Cock, K.) 87–100 (Wiley, Chichester, 1998).
12. Ghebreyesus, T. A. *et al.* Incidence of malaria among children living near dams in northern Ethiopia; community based incidence survey. *Brit. Med. J.* **319**, 663–666 (1999).
13. Lindblade, K. A., Walker, E. D., Onapa, A. W., Katungu, J. & Wilson, M. L. Highland malaria in Uganda: prospective analysis of an epidemic associated with El Nino. *Trans. R. Soc. Trop. Med. Hyg.* **93**, 480–487 (1999).
14. Brown, V., Issak, M. A., Rossi, M., Barboza, P. & Paugram, A. Epidemic of malaria in north-eastern Kenya. *Lancet* **352**, 1356–1357 (1998).
15. Muentener, P., Schlagenhauf, P. & Steffen, R. Imported malaria (1985–95): trends and perspectives. *Bull. World Health Org.* **77**, 560–566 (1999).
16. Martens, P. & Hall, L. Malaria on the move: human population movement and malaria transmission. *Emerg. Infect. Dis.* **6**, 103–109 (2000).
17. Roberts, D. R., Manguin, S. & Mouchet, J. DDT house spraying and re-emerging malaria. *Lancet* **356**, 330–332 (2000).
18. Lengeler, C. Insecticide treated bednets and curtains for malaria control. Cochrane Review, The Cochrane Library, issue no. 3 <<http://www.update-software.com/cochrane/abstracts/ab000363.htm>> (Update Software, Oxford, 1998).
19. Armstrong Schellenberg, J. R. M. *et al.* Effect of large-scale social marketing of insecticide-treated nets on child survival in rural Tanzania. *Lancet* **357**, 1241–1247 (2001).
20. Schultz, L. J. *et al.* The efficacy of antimalarial regimens containing sulfadoxine-pyrimethamine and/or chloroquine in preventing peripheral and placental *Plasmodium falciparum* infection among pregnant women in Malawi. *Am. J. Trop. Med. Hyg.* **51**, 515–522 (1994).
21. Shulman, C. E. *et al.* Intermittent sulphadoxine-pyrimethamine to prevent severe anaemia secondary to malaria in pregnancy: a randomised placebo-controlled trial. *Lancet* **353**, 632–636 (1999).
22. Parise, M. *et al.* Efficacy of sulfadoxine-pyrimethamine for prevention of placental malaria in an area of Kenya with a high prevalence of malaria and human immunodeficiency virus infection. *Am. J. Trop. Med. Hyg.* **59**, 813–822 (1998).
23. Schellenberg, D. *et al.* Intermittent treatment for malaria and anaemia control at time of routine vaccinations in Tanzanian infants: a randomised, placebo-controlled trial. *Lancet* **357**, 1471–1477 (2001).
24. Kidane, G. & Morrow, R. H. Teaching mothers to provide home treatment of malaria in Tigray, Ethiopia: a randomised trial. *Lancet* **356**, 550–555 (2000).
25. Marsh, V. M. *et al.* Changing home treatment of childhood fevers by training shop keepers in rural Kenya. *Trop. Med. Int. Health* **4**, 383–389 (1999).
26. Yeboah-Antwi, K. *et al.* Impact of prepackaging antimalarial drugs on cost to patients and compliance with treatment. *Bull. World Health Org.* **79**, 394–399 (2001).
27. Bloland, P. B., Ettling, M. & Meek, S. Combination therapy for Africa: hype or hope? *Bull. World Health Org.* **78**, 1378–1388 (2000).

The pathogenic basis of malaria

Louis H. Miller*, Dror I. Baruch*, Kevin Marsh† & Ogobara K. Doumbo‡

*Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

†KEMRI-Wellcome Trust Collaboration Programme, Center for Geographic Medicine Research Coast, Post Office Box 230, Kilifi, Kenya (e-mail: kmarsh@kilifi.mimcom.net)

‡Malaria Research and Training Center, Bamako, Mali (e-mail: okd@mrtcbko.malinet.ml)

Malaria is today a disease of poverty and underdeveloped countries. In Africa, mortality remains high because there is limited access to treatment in the villages. We should follow in Pasteur's footsteps by using basic research to develop better tools for the control and cure of malaria. Insight into the complexity of malaria pathogenesis is vital for understanding the disease and will provide a major step towards controlling it. Those of us who work on pathogenesis must widen our approach and think in terms of new tools such as vaccines to reduce disease. The inability of many countries to fund expensive campaigns and antimalarial treatment requires these tools to be highly effective and affordable.

Millions of children die from malaria in Africa every year¹. But the clinical outcome of an infection in a child depends on many factors (Fig. 1). These factors, often ill-defined, determine the outcome in each child. The top priority must be disease prevention because of the inability of the mothers to access or afford optimal treatment, and the ever-evolving drug resistance. Prevention may be effected through vector control such as insecticide-treated bednets or through the development of antimalarial vaccines.

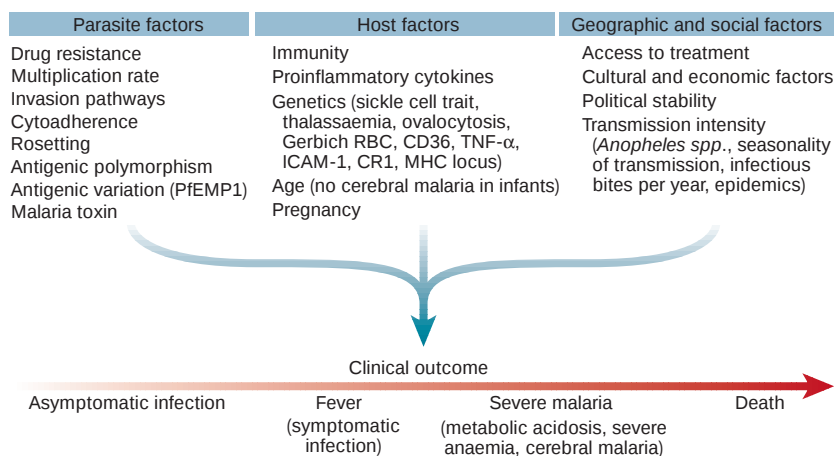
Malaria: the disease

Over the past 10 years, there have been several key shifts in our understanding of what constitutes severe malaria, and these shifts define the issues in pathogenesis that need to be explored to develop better treatments for sick children. The first shift is the increasing recognition that severe malaria is a disorder that affects several tissues and organs, even when the most marked manifestations may seem to involve a single organ such as the brain. In particular, metabolic acidosis, often profound, has been recognized as a principal

pathophysiological feature that cuts across the classical clinical syndromes of cerebral malaria and severe malarial anaemia². It is the single most important determinant of survival and leads directly to a common, but previously poorly recognized, syndrome of respiratory distress³. In most cases, this is predominantly (but not exclusively) a lactic acidosis⁴. There are several causes of lactic acidosis in children with severe malaria, from increased production of lactic acid by parasites (through direct stimulation by cytokines) to decreased clearance by the liver; however, most important by far is probably the combined effects of several factors that reduce oxygen delivery to tissues⁵.

A key feature of the biology of *Plasmodium falciparum* is its ability to cause infected red blood cells (RBCs) to adhere to the linings of small blood vessels. Such sequestered parasites cause considerable obstruction to tissue perfusion. In addition, in severe malaria there may be marked reductions in the deformability of uninfected RBCs^{6,7}. The pathogenesis of this abnormality is not clear, but its strong correlation with acidosis suggests that it may be involved in compromising blood flow through tissues. Individuals affected with malaria are often dehydrated and relatively hypovolaemic⁸,

Figure 1 The clinical outcome of malarial infection in an African child depends on many parasite, host, geographic and social factors. These converge in the child to result in a range of outcomes, from an asymptomatic infection to severe disease and death.



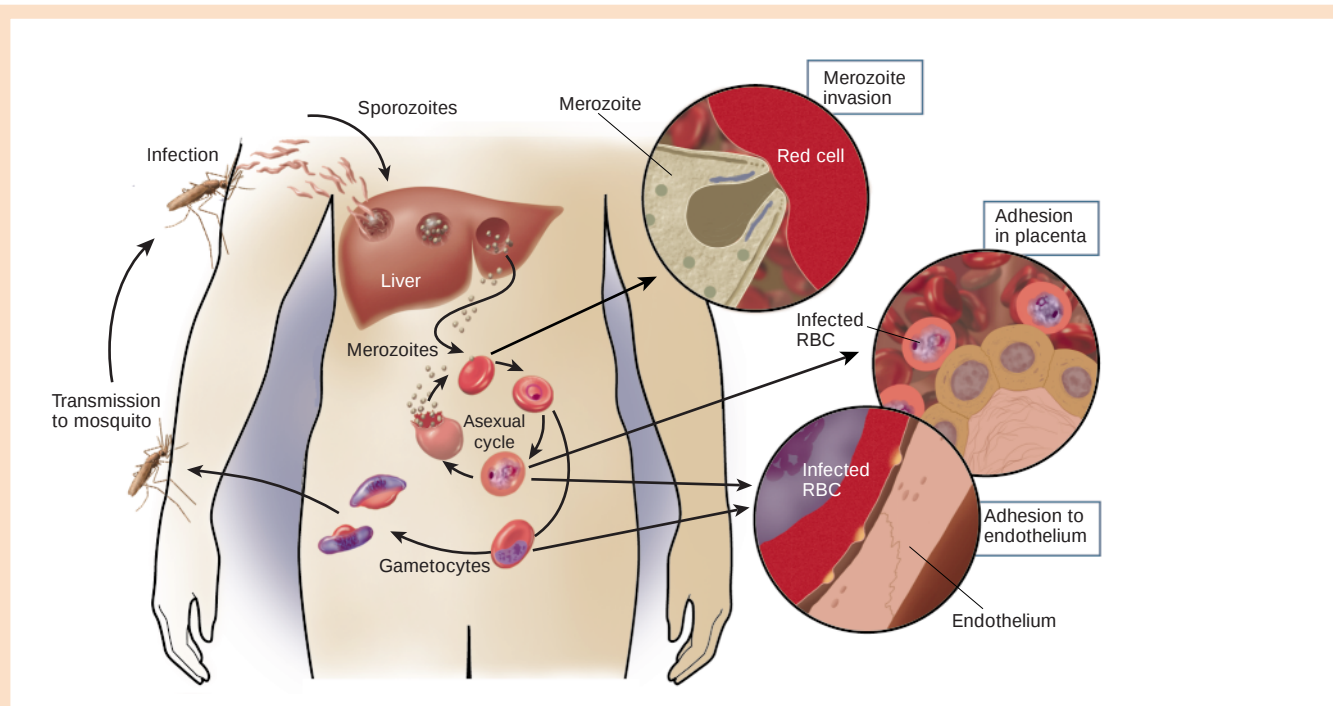


Figure 2 Parasite life cycle and pathogenesis of *falciparum* malaria. The molecular and cellular events during the parasite life cycle influence the severity of the disease. Disease occurs only as a result of the asexual blood stage after the parasite leaves the liver and begins to invade and grow inside red blood cells (RBCs). All human *Plasmodium* spp. invade by the same mechanism, but *P. falciparum* reaches high

parasitaemia because of greater flexibility in the receptor pathways that it can use to invade all RBCs. RBCs infected with *P. falciparum* must bind to endothelium or placenta for the parasite to avoid spleen-dependent killing mechanisms, but this binding also leads to much of the pathology (see Fig. 4).

which potentially exacerbates microvascular obstruction by reducing perfusion pressure. The destruction of RBCs is also an inevitable part of malaria, and anaemia further compromises oxygen delivery.

The second and related shift in our concept of severe malaria is the realization that there is no simple one-to-one correlation between the clinical syndromes and the pathogenic processes. Thus, severe anaemia may arise from many poorly understood mechanisms including acute haemolysis of uninfected RBCs and dyserythropoiesis, as well as through the interaction of malarial infection with other parasite infections and with nutritional deficiencies⁹. For many desperately sick children a simple 'one pathogen/one disease' model is not adequate, as bacteraemia caused by common pathogens may be present with acute malaria and may be a factor in mortality^{10,11}. Even the rigorously defined syndrome of cerebral malaria is used to describe children who have arrived at the point of coma through different routes. In many of these children, coma seems to be a response to overwhelming metabolic stress rather than a primary problem in the brain. Such children are often profoundly acidotic and may regain consciousness remarkably quickly after appropriate resuscitation¹², suggesting that cerebral malaria in this instance cannot be a consequence of the classical histologic picture.

Similarly, it has been recognized that a significant proportion of children in coma are, in fact, experiencing covert status epilepticus¹³, which responds rapidly to appropriate anticonvulsant therapy. The pathogenesis of this condition is unknown, but again the speed of resolution argues against classical views of pathogenesis. The picture that emerges is one in which many processes lead to a common outcome. These distinctions are much more than academic: they have direct implications for therapy, and they also identify the research issues needed to improve therapy for sick children.

Severe malaria is complex and probably cannot be represented accurately by any single scheme; however, our current understanding of the way in which several key pathogenic processes combine to cause severe disease invokes several basic processes: rapid expansion

of infected RBC mass, destruction of both infected and uninfected RBCs, microvascular obstruction, and inflammatory processes that combine to lead to reduced tissue perfusion. This, in turn, may lead to downstream events at a cellular level that further exacerbate the situation.

These general processes, which affect many tissue beds, may also be focused on specific organs in some situations, for instance the brain in cerebral malaria or the placenta during malaria in pregnancy. This could reflect both host-specific factors (for example, an increased likelihood to express particular receptors on cerebral endothelium) and parasite-specific factors (for example, the expression of molecules on the infected RBCs surface that are particularly suited for binding to certain receptors). In this article, we review the main advances in our understanding of malaria pathogenesis with the hope that these advances will lead to new tools to prevent disease before children become so sick that they need hospitalization.

Although the disease must ultimately be understood in humans, much of our knowledge of pathogenesis depends on studies in non-human species and *in vitro* cultures of *P. falciparum*. The parasitic invasion of hepatocytes and RBCs studied in rodent malarias caused by *Plasmodium berghei* and *Plasmodium yoelii*, and the rhesus malaria caused by *Plasmodium knowlesi*, respectively, have provided insight into these processes. Inflammatory cytokines are often studied in rodent malarias. In addition, these parasite species are important for the screening of drugs and vaccines, including those targeted at human malarias, in New World primates.

***Plasmodium* life cycle and pathogenesis**

Plasmodium falciparum and, to a much lesser extent, *Plasmodium vivax*¹⁴ are the main causes of disease and death from malaria. Mosquitoes inject parasites (sporozoites) into the subcutaneous tissue, and less-frequently directly into the bloodstream; from there, sporozoites travel to the liver (Fig. 2). Evidence indicates that sporozoites pass through several hepatocytes before invasion is followed by

parasite development¹⁵. The co-receptor on sporozoites that mediates invasion involves, in part, the thrombospondin domains on the circumsporozoite protein and on thrombospondin-related adhesive protein (TRAP). These domains bind specifically to heparin sulphate proteoglycans on hepatocytes in the region in apposition to sinusoidal endothelium and Kupffer cells¹⁶. Inside the hepatocyte, each sporozoite develops into tens of thousands of merozoites, which can each invade an RBC on release from the liver. Disease begins only once the asexual parasite multiplies in RBCs. This is the only gateway to disease.

Plasmodium falciparum and *P. vivax* develop over 48 hours in RBCs, producing around 20 merozoites per mature parasite, with each merozoite able to invade other RBCs. A small proportion of asexual parasites converts to gametocytes that are essential for transmitting the infection to others through female anopheline mosquitoes, but cause no disease. Here, the strategy of *P. vivax* differs from that of *P. falciparum*. *P. vivax* develops into gametocytes soon after the release of merozoites from the liver; *P. falciparum* gametocytes develop much later. The early treatment of clinical attacks of malaria by anti-bloodstage chemotherapy for *P. falciparum* also kills the developing gametocytes; *P. vivax* transmits before the symptomatic stage of the disease.

Invasion of RBCs

The sequence of invasion is probably similar for all *Plasmodium* spp. The parasite must engage binding receptors¹⁷ on the RBC, and undergo apical reorientation¹⁸, junction formation¹⁹, and signalling. The parasite induces a vacuole derived from the RBC's plasma membrane and enters the vacuole by a moving junction. Three organelles on the invasive (apical) end of the parasite (rhoptries, micronemes and dense granules) define the phylum Apicomplexa. Receptors that mediate invasion of RBCs by merozoites and invasion of liver by sporozoites are found in micronemes²⁰, on the cell surface, and in rhoptries. The location of these receptors within organelles may protect the parasite from antibody-mediated neutralization, as the release from apical organelles after contact with the RBC may limit their exposure to antibody.

Identifying the signalling pathways that release organelle contents on contact with a host RBC is a critical issue in parasite biology (Fig. 2). Malaria parasites have intracellular signalling pathways mediated by phosphoinositide, cyclic AMP and calcium-dependent mechanisms. What remains completely unknown is which mero-

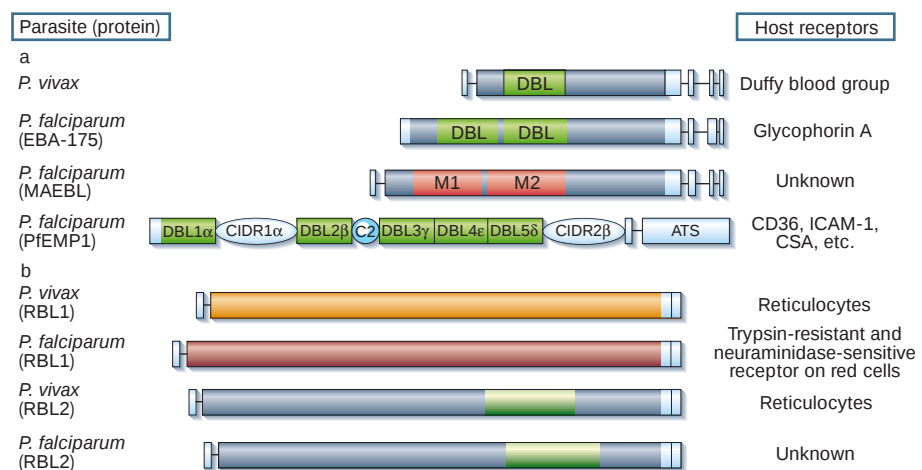
zoite surface molecules recognize the RBC surface and then signal the start of the invasion process. Invasion events include releasing essential molecules from apical organelles and initiating the actin-myosin moving junction that brings the parasite inside the vacuole that forms in the RBC. The TRAP protein interacts with skeletal proteins in malaria sporozoites and in *Toxoplasma gondii*²¹, but the equivalent molecule for merozoites has yet to be identified.

Both *P. falciparum* and *P. vivax* can cause severe anaemia, but only *P. falciparum* causes the many complications of cerebral malaria, hypoglycaemia, metabolic acidosis and respiratory distress. Certain differences in the biology of the two parasites partially explain the differences in patterns of disease. First, *P. falciparum* can invade a large percentage of RBCs, whereas *P. vivax* is limited to reticulocytes. Similar differences are found between virulent and avirulent *P. yoelii*. Both invade reticulocytes preferentially, but once the reticulocytes are consumed virulent *P. yoelii* can invade all RBCs, leading to higher parasitaemia and death. A comparison between severe and uncomplicated falciparum malaria has suggested a similar pattern of RBC invasion, with the virulent *P. falciparum* invading all RBCs and the avirulent parasites invading only a subpopulation²².

A second difference is the surprising redundancy of invasion pathways in *P. falciparum* as compared with *P. vivax*. The latter parasite invades only Duffy blood group positive RBCs²³ and is largely limited to reticulocytes. In West Africa, where RBCs are Duffy blood group negative, *P. vivax* has essentially disappeared. The Duffy negative blood group has arisen independently in Papua New Guinea²⁴ — a region in which *P. vivax* is highly endemic. The limitations of *P. vivax* invasion have led to the discovery of two families of parasite receptors (Fig. 3): first, the parasite molecule that binds to the Duffy blood group system, and homologous Duffy-binding-like (DBL) proteins of *P. falciparum* and *P. knowlesi*²⁵; and second, the parasite reticulocyte-binding proteins of *P. vivax*²⁶, and homologous reticulocyte-binding-like (RBL) proteins of *P. falciparum*²⁷ and *P. yoelii*²⁸. The various members of the DBL and RBL families may recognize different RBC receptors to those of the Duffy blood group or the receptor on reticulocytes. The receptor grouping into DBL and RBL refers to the family of homologous parasite proteins, not the binding specificity on the RBC.

Plasmodium yoelii possesses a large family of RBL genes. Each of the merozoites in a single infected RBC can express a different member of the RBL family²⁹. If each has a different RBC-binding specificity, then the parasite has a greater chance for survival. Thus,

Figure 3 Two families of *Plasmodium* spp. receptors. **a**, The Duffy-binding-like (DBL) family is named after the receptor region of *P. vivax* that binds the Duffy blood group protein on the surface of RBCs. Homologous proteins with a DBL receptor region occur in many *Plasmodium* spp. Whereas *P. vivax* possesses a single DBL protein, *P. falciparum* has several that use RBC receptors other than Duffy, and may in part explain the ability of the parasite to invade all aged RBCs. MAEBL has the same structure as the *P. vivax* protein except that its receptor domain has been replaced by a region that is homologous to another parasite protein. In addition to its function in invasion, DBL is also important as a receptor domain on the *P. falciparum* erythrocyte membrane protein 1 (PfEMP1), which is expressed on the infected RBC surface. See Fig. 4 for the role of PfEMP1 in pathology. **b**, A second family, the reticulocyte binding-like (RBL) parasite proteins, was first discovered in *P. vivax* through its binding to reticulocytes. In *P. vivax*, RBL is present as a RBL1/RBL2 heterodimer.



RBL2 has a defined highly homologous region (shaded in green). The receptor domains of the heterodimer remain to be defined, as does the requirement of the heterodimer for binding RBCs in *Plasmodium* spp. other than *P. vivax*.

although the full details of the DBL and RBL families are unknown, these receptors clearly determine much of the flexibility for invasion by the various *Plasmodium* spp. This flexibility, in turn, determines the maximum parasitaemia and disease caused by the various parasites.

Plasmodium falciparum can use its many redundant pathways to invade at equal or reduced efficiency RBCs that lack a particular receptor such as sialic acid^{30,31}. Three sialoglycoprotein-dependent pathways involving RBCs and parasite co-receptors have been identified: glycophorin A and the parasite DBL protein EBA-175 (ref. 32); glycophorin C/D and the DBL parasite protein BAEBL³³; and a trypsin-resistant pathway involving a *P. falciparum* RBL protein²⁷. A fourth may involve sialic acid on glycophorin B (ref. 27).

Despite markedly reduced invasion of glycophorin-A-negative RBCs, only glycophorin B mutations occur in Africa. Gerbich RBCs do not express glycophorin D, express an altered glycophorin C and have reduced binding to the parasite molecule BAEBL. Gerbich RBCs are rare in most parts of the world except in the falciparum-endemic regions of Papua New Guinea, where the allele frequency approaches 50% (ref. 34). Such redundancy and alternative pathways are a large advantage to the survival of *P. falciparum* in response to changes in host genetics. The parasite, however, may become less virulent as it adapts to survival in these deficient RBCs.

Studying the DBL and RBL families has begun to yield a molecular understanding of the diverse invasion pathways for *P. falciparum* and other *Plasmodium* spp. Although other parasite proteins on the merozoite surface and in apical organelles have been proposed as receptors^{35–37}, there is no direct evidence so far. Because invasion is such a complex series of events from RBC binding, to apical reorientation, to entry, it seems likely that several proteins are required for efficient invasion. For example, evidence has suggested that RBC invasion requires the cleavage of a surface protein on the RBC by a parasite serine protease³⁸. This parasite enzyme has yet to be identified. Thus, the molecular and cellular events surrounding each step in invasion still remain to be elucidated. Understanding these pathways will give insight into parasite virulence and will facilitate rational vaccine design against merozoite invasion.

Binding of parasitized RBCs to vascular endothelium and placenta

An important difference between *P. falciparum* and other human malarias is the way in which *P. falciparum* modifies the surface of the RBCs so that asexual parasites and gametocytes can adhere to the endothelium and asexual parasites can adhere to the placenta. As a result, only ring forms of *P. falciparum* are found in circulating blood (for review, see refs 39–41). The surface of *P. falciparum* trophozoite- and schizont-infected RBCs is covered with knob-like excrescences that are the contact points with host cells⁴². Adherence protects the parasite from destruction, as non-adherent mature parasitized RBCs are cleared rapidly in the spleen⁴³.

Attempts to decipher the highly complex and pathogenic adhesion process emphasize how much we have learned and how little we understand. To determine whether and how sequestration can lead to pathogenesis, we should first look at how the parasite sequesters. The *P. falciparum* adhesion process, in which most parasites first tether and then roll, before becoming firmly adherent^{44,45}, is comparable to leukocyte adhesion. Most host receptors are involved with tethering and rolling but are unable to support on their own firm adhesion under flow^{44,46}. Binding to these host receptors is important, however, as it significantly increases adhesion, which may allow the parasite to bind efficiently to the endothelium of various organs⁴⁷. Only two receptors, CD36 and chondroitin sulphate A (CSA), provide stable stationary adherence^{44,46}.

Parasites sequester themselves in various organs including heart, lung, brain, liver, kidney, subcutaneous tissues and placenta. The various endothelial cells in these organs and syncytiotrophoblasts in placenta express different and variable amounts of host receptors. To successfully adhere to these cells, the parasite can bind to several

receptors (ref. 39; and Fig. 4). The adhesion phenotype is not homogenous, and different parasites can bind to variable numbers and combinations of host receptors^{48,49}. This variability is believed to affect the tissue distribution and pathogenesis of parasites.

A single parasite protein — *P. falciparum* erythrocyte membrane protein 1 (PfEMP1), which is expressed at the infected erythrocyte surface — mediates parasite binding to all the various receptors^{39–41}. PfEMP1 is encoded by the large and diverse *var* gene family that is involved in clonal antigenic variation and has a central role in *P. falciparum* pathogenesis^{50–52}. The multiple adhesion domains located at the extracellular region of PfEMP1 can simultaneously recognize several host receptors. These domains contain variable numbers of different (five types) DBL domains, named for their homology to the DBL domains involved in RBC invasion, and 1–2 cysteine-rich interdomain regions (CIDRs)^{52,53}. The binding domains for several host receptors have been mapped to various DBL and CIDR domains^{39,54}. The diversity in this gene family is extensive, and numerous *var* genes appear in the parasite population. Although each parasite within an RBC expresses a single *var* gene⁵⁵, other *var* genes in its repertoire (out of 50 in its genome) can be expressed up to a rate of 2% per parasite growth cycle⁵⁶.

In most cases, the binding to host endothelium does not lead to pathogenesis, as most infections result in malaria that is devoid of complications⁵⁷. What causes the transition from an uncomplicated to a serious infection, such as cerebral malaria, is unclear at present. An intriguing possibility is that the expression of particular binding properties will lead to distinct patterns of sequestration and to pathogenic consequences. One example is the sequestration of infected RBCs in the placenta, which causes premature delivery, low birth weight and increased mortality in the newborn, as well as anaemia in the mother. Parasitized RBCs isolated from placentas have a unique adhesion property that is different from parasites collected from non-pregnant individuals^{48,58}. These parasites bind to CSA but fail to adhere to CD36 — the crucial host receptor for sequestration in microvasculature. The apparent dichotomy in adhesion to these receptors has been selected to allow the parasite to sequester not in endothelium but in placenta — perhaps a site of reduced immunity. Indeed, CSA-binding parasites express PfEMP1 with a DBL- γ domain that binds CSA and a non-CD36-binding CIDR1 (refs 59,60). In contrast, CD36-adherent parasites express a PfEMP1 with a CD36-binding CIDR1 (ref. 60).

Sequestration of parasites in the brain may be related to cerebral malaria and may involve the intercellular adhesion molecule 1 (ICAM-1) receptor⁴¹. Although infected RBCs are bound to brain endothelium at autopsy, it is unknown whether this represents a different distribution of adhesion from that of uncomplicated malaria. An increase in the expression of ICAM-1 in brain endothelium may explain differences in parasite adhesion in cerebral malaria^{61,62}. The role of sequestration in other severe complications of malaria remains unclear. Pathogenic connections between adhesion and host receptors are supported by both a nonsense mutation in the gene of the adhesion receptor CD36 that is associated with protection from severe malaria, and the link between complement receptor 1 and ABO blood group antigens and rosetting (the binding of uninfected RBCs)^{40,63,64}. Several investigators have suggested that simultaneous binding to multiple receptors might be associated with more severe cases of malaria⁶⁵, but specific data are lacking. Some properties, such as rosetting⁴⁰ and clumping⁶⁶, appear at higher frequencies in cases of severe malaria, but these associations have not been found in all studies and their effect on pathogenesis remains obscure. One possibility is that competition (for adhesion) between parasites drives some of them to develop new adhesion properties and sequester themselves in less desirable locations that lead to pathogenesis.

Although dissecting various individual interactions is a good experimental approach, the outcome of an infection and progression into pathology depend on the specific and dynamic combination of the host and the parasite properties. Clinical disease also changes

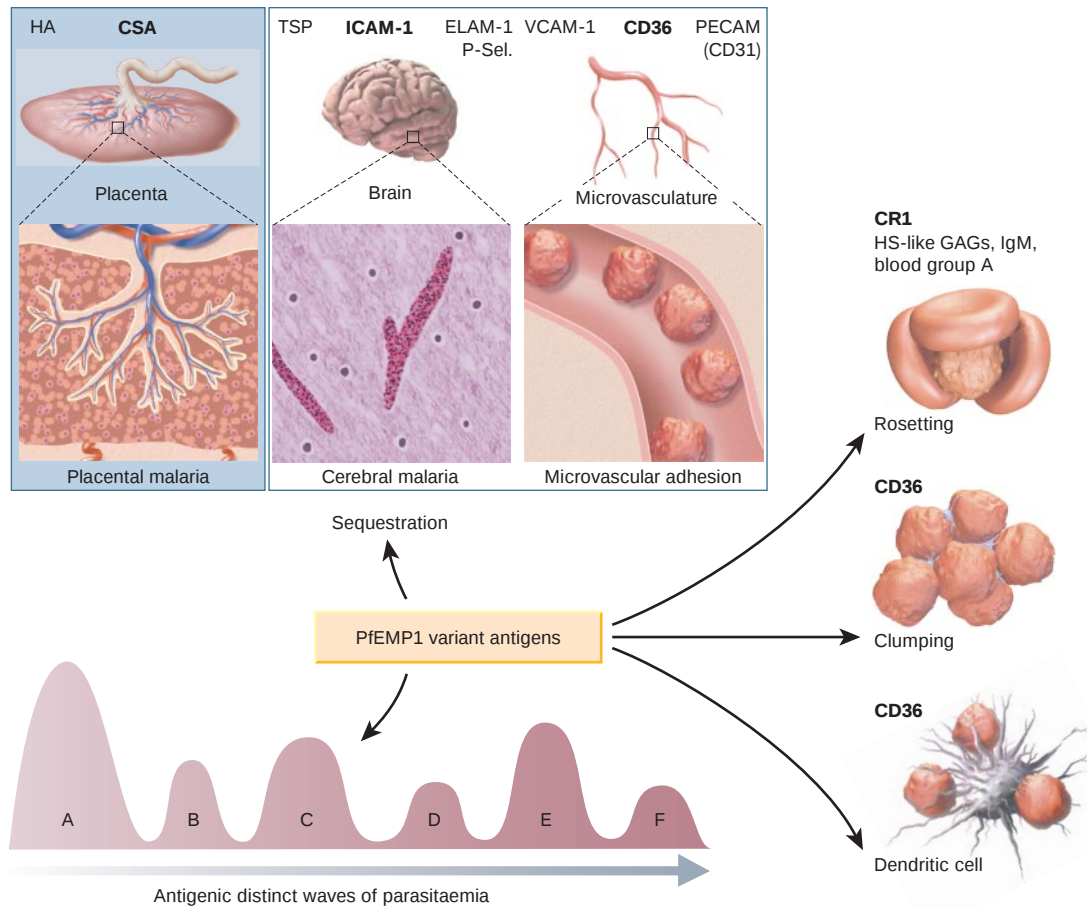


Figure 4 The variant antigen family of PfEMP1 is central to host–parasite interaction and pathogenesis. PfEMP1 expressed on the surface of mature RBCs infected with *P. falciparum* is involved with clonal antigenic variation and can bind to many host receptors through its multiple adhesion domains (Fig. 3). The different properties of PfEMP1 — sequestration for evading spleen-dependent killing and antigenic variation for evading antibody-dependent killing — contribute to the virulence and pathogenesis of *P. falciparum* and are essential for survival of the parasite. Parasite sequestration in the brain and placenta contribute to the complications of cerebral malaria and placental malaria, respectively. Simultaneous binding to several receptors, binding of uninfected

erythrocytes (rosetting), and clumping of infected erythrocytes through platelets are associated with the pathogenesis of malaria. Parasite-infected RBCs binding to dendritic cells downregulates the host immune response. HA, hyaluronic acid; TSP, thrombospondin; ELAM-1, endothelial/leukocyte adhesion molecule 1; P-Sel., P-selectin; VCAM-1, vascular cell adhesion molecule 1; PECAM (CD31), platelet endothelial cell adhesion molecule 1; CR1, complement receptor 1; HS-like GAGs, heparin sulphate-like glycosaminoglycans; IgM, immunoglobulin M. Other abbreviations defined in the text.

with age, immunity and transmission rates⁵⁷. Immunity to malaria has a major role in controlling disease and pathogenesis. The properties of PfEMP1 as an adhesion protein (to avoid parasite destruction in the spleen) cannot be separated from its involvement in immune evasion by clonal antigenic variation, which can lead to chronic infection. Even after many exposures, humans are not refractory to malaria parasites but develop clinical immunity that prevents symptomatic disease. This type of immunity limits disease and, although the individual may carry low numbers of parasites, they do not develop into a symptomatic infection⁵⁷. The role of anti-PfEMP1 antibodies in protecting against pathogenic infections is highlighted again in placental malaria. Exposure to parasites that sequester themselves in the placenta during pregnancy induces strain-transcending immunity that stops infected erythrocytes adhering to CSA and may protect the mother and fetus from placental malaria in subsequent pregnancies⁶⁷.

During the development of clinical immunity, particularly during early childhood, strain-specific antibodies to PfEMP1 are important in preventing infection with previously encountered

isolates^{68,69}. This protection can be significant during and after infections with virulent isolates. Bull *et al.*^{69,70} have provided evidence for rare and prevalent isolates, and shown that parasites causing severe disease tend to express a subset of variant surface antigens (PfEMP1). Moreover, these isolates were expressed preferentially in children who were less able to recognize (by antibody-mediated agglutination) many isolates. Children exposed once or twice to non-cerebral severe malaria acquire immunity that protects them from this form of the disease⁷¹. Hence, exposure to pathogenic forms of *P. falciparum* can protect against these parasites, leading to the selection of possibly less virulent parasites in subsequent infections. Despite its variation, regions of PfEMP1 are restricted by function (for example, binding to CD36 or CSA), and these regions may be potential targets for vaccines.

How adhesion progresses to pathology is a principal issue that remains unresolved. Several mechanisms that might cause damage to host endothelium and organs have been proposed, including obstruction of blood flow, and systemic or local production and deposition of pro-inflammatory cytokines (see below). Parasite

adhesion can also affect the endothelium by inducing or blocking signal transduction mediated by host receptors such as CD36. Advances in adhesion research will hopefully provide clues to the mechanism underlying adhesion-related pathogenesis.

The pro-inflammatory immune response and pathogenesis

Antibodies and the pro-inflammatory response protect against the asexual blood stages of some rodent malarias and probably also human malaria. Protection mediated by the pro-inflammatory response may relate to the cytokines tumour-necrosis factor- α (TNF- α) and interferon- γ (IFN- γ), and the release of mediators such as nitric oxide (NO). Clark and Cowden⁷² have proposed that mediators, especially NO, are also central to disease. Although it is perfectly logical that these are involved in bone marrow suppression and cerebral malaria, there are no data to prove this as yet. Furthermore, we do not have a animal model in which to study cerebral malaria. One hypothesis suggests that TNF- α induces brain endothelial cells to express ICAM-1 (ref. 73), as vessels in the brain have increased expression of ICAM-1 in cerebral malaria⁶². Although NO has been proposed as the cause of cerebral malaria, there are higher systemic concentrations of NO in uncomplicated malaria than in cerebral malaria⁷⁴. Coma might be caused by local increases in NO in the brain and not by increased concentrations in blood; however, this has not been measured. Indeed, total nitrate plus nitrite levels in the cerebrospinal fluid of children with cerebral malaria are low, and it has been suggested that this may exacerbate NMDA (N-methyl-D-aspartate)-mediated neurotoxicity caused by excitotoxins such as quinolinic acid⁷⁵.

Data that suggest that a toxin of malarial origin drives the pro-inflammatory response are interesting^{76,77}, but the physiological significance of this remains to be proved. Evidence that a particular molecule is involved in inducing the pro-inflammatory response has come from an assay measuring the release of TNF- α by macrophages *in vitro*. The isolation of subcellular components from the parasite coupled with this *in vitro* assay led to the identification of the glycosylphosphatidylinositol (GPI) anchor from parasite proteins MSP1 and MSP2 as an inducer of pro-inflammatory cytokines. Antibodies to the GPI anchor are associated with a lack of disease in adults⁷⁸, but there is no proof that this is causally related.

Modifications in the immune response to malaria that may not be specific to this disease have been identified. Infection with *P. falciparum* causes apoptosis of mononuclear cells in infected humans⁷⁹. Infecting mice with a rodent malaria to which they had been previously exposed led to the apoptosis of T cells immune to malaria and not those immune to ova, a malaria-unrelated antigen⁸⁰. The cells that were eliminated were pro-inflammatory T cells, which produce IFN- γ and interleukin (IL)-2, but not IL-4. It is unclear whether this is specific to malaria or a more general phenomenon.

Studying genetic differences between populations may inform our understanding of the immune system. Fulanis, an ethnic group in Burkina Faso, have a higher titre of antibodies to many malarial antigens and less disease than two other ethnic groups in the same village who are bitten by equal numbers of infected mosquitoes⁸¹. The molecular basis is unknown, but the innate immune system may be interacting with the adaptive system to increase antibody titres. The importance of these differences is also highlighted by a study of insecticide-impregnated bednets⁸². The use of insecticide-treated bednets has reduced the infection in Fulanis, but not among other ethnic groups in the same area. Possibly, the higher antibody titres in Fulanis are sufficient to take advantage of the reduction in infectious bites as a result of insecticide-treated bednets.

Perspectives

The clinical outcome of an infection in a child in Africa depends on many factors (Fig. 1). In our attempt to understand disease, we often take a reductionist view and study individual components of the parasite and human in an attempt to identify factors that have a large

impact on disease outcome. Such factors can be targets for intervention through the development of new tools such as vaccines. Success in the development and implementation of these new tools will depend on a connection with scientists from endemic countries of Africa who have a better understanding of local customs and are experienced in communicating with the poorest people in villages of Africa. □

1. Snow, R. W., Craig, M., Deichmann, U. & Marsh, K. Estimating mortality, morbidity and disability due to malaria among Africa's non-pregnant population. *Bull. World Health Organ.* **77**, 624–640 (1999).
2. Marsh, K. *et al.* Indicators of life-threatening malaria in African children. *N. Engl. J. Med.* **332**, 1399–1404 (1995).
3. Taylor, T. E., Borgstein, A. & Molyneux, M. E. Acid–base status in paediatric *Plasmodium falciparum* malaria. *Q. J. Med.* **86**, 99–109 (1993).
4. English, M. *et al.* Deep breathing in children with severe malaria: indicator of metabolic acidosis and poor outcome. *Am. J. Trop. Med. Hyg.* **55**, 521–524 (1996).
5. English, M. *et al.* Acidosis in severe childhood malaria. *Q. J. Med.* **90**, 263–270 (1997).
6. Miller, L. H., Shunichi, U. & Chien, S. Alteration in the rheologic properties of *Plasmodium knowlesi* infected red cells. A possible mechanism of cerebral malaria. *J. Clin. Invest.* **50**, 1451–1455 (1971).
7. Dondorp, A. M., Kager, P. A., Vreeken, J. & White, N. J. Abnormal blood flow and red blood cell deformability in severe malaria. *Parasitol. Today* **16**, 228–232 (2000).
8. English, M. C. *et al.* Hyponatraemia and dehydration in severe malaria. *Arch. Dis. Child.* **74**, 201–205 (1996).
9. Newton, C. R. *et al.* Severe anaemia in children living in a malaria endemic area of Kenya. *Trop. Med. Int. Health* **2**, 165–178 (1997).
10. Berkley, J., Mwarumba, S., Bramham, K., Lowe, B. & Marsh, K. Bacteraemia complicating severe malaria in children. *Trans. R. Soc. Trop. Med. Hyg.* **93**, 283–286 (1999).
11. Prada, J., Alabi, S. A. & Bienzie, U. Bacterial strains isolated from blood cultures of Nigerian children with cerebral malaria. *Lancet* **342**, 1114 (1993).
12. English, M., Waruiru, C. & Marsh, K. Transfusion for respiratory distress in life-threatening childhood malaria. *Am. J. Trop. Med. Hyg.* **55**, 525–530 (1996).
13. Crawley, J. *et al.* Seizures and status epilepticus in childhood cerebral malaria. *Q. J. Med.* **89**, 591–597 (1996).
14. Mendis, K., Sina, B. J., Marchesini, P. & Carter, R. The neglected burden of *Plasmodium vivax* malaria. *Am. J. Trop. Med. Hyg.* **64** (Suppl. 1–2), 97–105 (2001).
15. Mota, M. M. *et al.* Migration of *Plasmodium sporozoites* through cells before infection. *Science* **291**, 141–144 (2001).
16. Frevet, U. *et al.* Malaria circumsporozoite protein binds to heparan sulfate proteoglycans associated with the surface membrane of hepatocytes. *J. Exp. Med.* **177**, 1287–1298 (1993).
17. Chittis, C. E. Molecular insights into receptors used by malaria parasites for erythrocyte invasion. *Curr. Opin. Hematol.* **8**, 85–91 (2001).
18. Dvorak, J. A., Miller, L. H., Whitehouse, W. C. & Shiroishi, T. Invasion of erythrocytes by malaria merozoites. *Science* **187**, 748–750 (1975).
19. Aikawa, M., Miller, L. H., Johnson, J. & Rabbege, J. Erythrocyte entry by malarial parasites. A moving junction between erythrocyte and parasite. *J. Cell Biol.* **77**, 72–82 (1978).
20. Adams, J. H. *et al.* The Duffy receptor family of *Plasmodium knowlesi* is located within the micronemes of invasive malaria merozoites. *Cell* **63**, 141–153 (1990).
21. Kappe, S. *et al.* Conservation of a gliding motility and cell invasion machinery in Apicomplexan parasites. *J. Cell Biol.* **147**, 937–944 (1999).
22. Chotivanich, K. *et al.* Parasite multiplication potential and the severity of falciparum malaria. *J. Infect. Dis.* **181**, 1206–1209 (2000).
23. Miller, L. H., Mason, S. J., Clyde, D. F. & McGinniss, M. H. The resistance factor to *Plasmodium vivax* in blacks. The Duffy-blood-group genotype, FyFy. *N. Engl. J. Med.* **295**, 302–304 (1976).
24. Zimmerman, P. A. *et al.* Emergence of FY*A^u in a *Plasmodium vivax*-endemic region of Papua New Guinea. *Proc. Natl Acad. Sci. USA* **96**, 13973–13977 (1999).
25. Adams, J. H. *et al.* A family of erythrocyte binding proteins of malaria parasites. *Proc. Natl Acad. Sci. USA* **89**, 7085–7089 (1992).
26. Galinski, M. R., Medina, C. C., Ingravallo, P. & Barnwell, J. W. A reticulocyte-binding protein complex of *Plasmodium vivax* merozoites. *Cell* **69**, 1213–1226 (1992).
27. Rayner, J. C., Vargas-Serrato, E., Huber, C., Galinski, M. R. & Barnwell, J. W. A *Plasmodium falciparum* homologue of *Plasmodium vivax* reticulocyte binding protein (PvRBP1) defines a trypsin-resistant erythrocyte invasion pathway. *J. Exp. Med.* **194**, 1571–1582 (2001).
28. Keen, J. K., Sinha, K. A., Brown, K. N. & Holder, A. A. A gene coding for a high-molecular mass rhoptry protein of *Plasmodium yoelii*. *Mol. Biochem. Parasitol.* **65**, 171–177 (1994).
29. Preisner, P. R., Jarra, W., Capiod, T. & Snounou, G. A rhoptry-protein-associated mechanism of clonal phenotypic variation in rodent malaria. *Nature* **398**, 618–622 (1999).
30. Okoyeh, J. N., Pillai, C. R. & Chittis, C. E. *Plasmodium falciparum* field isolates commonly use erythrocyte invasion pathways that are independent of sialic acid residues of glycophorin A. *Infect. Immun.* **67**, 5784–5791 (1999).
31. Dolan, S. A., Miller, L. H. & Wellems, T. E. Evidence for a switching mechanism in the invasion of erythrocytes by *Plasmodium falciparum*. *J. Clin. Invest.* **86**, 618–624 (1990).
32. Sim, B. K., Chittis, C. E., Wasniowska, K., Hadley, T. J. & Miller, L. H. Receptor and ligand domains for invasion of erythrocytes by *Plasmodium falciparum*. *Science* **264**, 1941–1944 (1994).
33. Mayer, D. C., Kaneko, O., Hudson-Taylor, D. E., Reid, M. E. & Miller, L. H. Characterization of a *Plasmodium falciparum* erythrocyte-binding protein paralogous to EBA-175. *Proc. Natl Acad. Sci. USA* **98**, 5222–5227 (2001).
34. Patel, S. S. *et al.* The association of the glycophorin C exon 3 deletion with ovalocytosis and malaria susceptibility in the Wosera, Papua New Guinea. *Blood* **98**, 3489–3491 (2001).
35. Patiño, J. A. G., Holder, A. A., McBride, J. S. & Blackman, M. J. Antibodies that inhibit malaria merozoite surface protein-1 processing and erythrocyte invasion are blocked by naturally acquired human antibodies. *J. Exp. Med.* **186**, 1689–1699 (1997).
36. Hodder, A. N. *et al.* The disulfide bond structure of *Plasmodium* apical membrane antigen-1. *J. Biol. Chem.* **271**, 29446–29452 (1996).

37. Triglia, T. *et al.* Apical membrane antigen 1 plays a central role in erythrocyte invasion by *Plasmodium* species. *Mol. Microbiol.* **38**, 706–718 (2000).
38. Braun-Breton, C. *et al.* *Plasmodium chabaudi* p68 serine protease activity required for merozoite entry into mouse erythrocytes. *Proc. Natl Acad. Sci. USA* **89**, 9647–9651 (1992).
39. Baruch, D. I. Adhesive receptors on malaria-parasitized red cells. *Baillieres Best Pract. Res. Clin. Haematol.* **12**, 747–761 (1999).
40. Chen, Q., Schlichtherle, M. & Wahlgren, M. Molecular aspects of severe malaria. *Clin. Microbiol. Rev.* **13**, 439–450 (2000).
41. Newbold, C. *et al.* Cytoadherence, pathogenesis and the infected red cell surface in *Plasmodium falciparum*. *Int. J. Parasitol.* **29**, 927–937 (1999).
42. Luse, S. A. & Miller, L. H. *Plasmodium falciparum* malaria. Ultrastructure of parasitized erythrocytes in cardiac vessels. *Am. J. Trop. Med. Hyg.* **20**, 655–660 (1971).
43. Langreth, S. G. & Peterson, E. Pathogenicity, stability, and immunogenicity of a knobless clone of *Plasmodium falciparum* in Colombian owl monkeys. *Infect. Immun.* **47**, 760–766 (1985).
44. Cooke, B. M. *et al.* Rolling and stationary cytoadhesion of red blood cells parasitized by *Plasmodium falciparum*: separate roles for ICAM-1, CD36 and thrombospondin. *Br. J. Haematol.* **87**, 162–170 (1994).
45. Ho, M. & White, N. J. Molecular mechanisms of cytoadherence in malaria. *Am. J. Physiol.* **276**, C1231–C1242 (1999).
46. Rogerson, S. J., Novakovic, S., Cooke, B. M. & Brown, G. V. *Plasmodium falciparum*-infected erythrocytes adhere to the proteoglycan thrombospondin in static and flow-based systems. *Exp. Parasitol.* **86**, 8–18 (1997).
47. Yip, B. G. *et al.* Synergism of multiple adhesion molecules in mediating cytoadherence of *Plasmodium falciparum*-infected erythrocytes to microvascular endothelial cells under flow. *Blood* **96**, 2292–2298 (2000).
48. Beeson, J. G. *et al.* *Plasmodium falciparum* isolates from infected pregnant women and children are associated with distinct adhesive and antigenic properties. *J. Infect. Dis.* **180**, 464–472 (1999).
49. Newbold, C. *et al.* Receptor-specific adhesion and clinical disease in *Plasmodium falciparum*. *Am. J. Trop. Med. Hyg.* **57**, 389–398 (1997).
50. Baruch, D. I. *et al.* Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* **82**, 77–87 (1995).
51. Smith, J. D. *et al.* Switches in expression of *Plasmodium falciparum* var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82**, 101–110 (1995).
52. Su, X. Z. *et al.* The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell* **82**, 89–100 (1995).
53. Smith, J. D., Subramanian, G., Gamain, B., Baruch, D. I. & Miller, L. H. Classification of adhesive domains in the *Plasmodium falciparum* erythrocyte membrane protein 1 family. *Mol. Biochem. Parasitol.* **110**, 293–310 (2000).
54. Smith, J. D., Gamain, B., Baruch, D. I. & Kyes, S. Decoding the language of var genes and *Plasmodium falciparum* sequestration. *Trends Parasitol.* **17**, 538–545 (2001).
55. Chen, Q. *et al.* Developmental selection of var gene expression in *Plasmodium falciparum*. *Nature* **394**, 392–395 (1998).
56. Roberts, D. J. *et al.* Rapid switching to multiple antigenic and adhesive phenotypes in malaria. *Nature* **357**, 689–692 (1992).
57. Snow, R. W. & Marsh, K. New insights into the epidemiology of malaria relevant for disease control. *Br. Med. Bull.* **54**, 293–309 (1998).
58. Fried, M. & Duffy, P. E. Adherence of *Plasmodium falciparum* to chondroitin sulfate A in the human placenta. *Science* **272**, 1502–1504 (1996).
59. Buffet, P. A. *et al.* *Plasmodium falciparum* domain mediating adhesion to chondroitin sulfate A: a receptor for human placental infection. *Proc. Natl Acad. Sci. USA* **96**, 12743–12748 (1999).
60. Gamain, B., Smith, J. D., Miller, L. H. & Baruch, D. I. Modifications in the CD36 binding domain of the *Plasmodium falciparum* variant antigen are responsible for the inability of chondroitin sulfate A adherent parasites to bind CD36. *Blood* **97**, 3268–3274 (2001).
61. Ockenhouse, C. F. *et al.* Molecular basis of sequestration in severe and uncomplicated *Plasmodium falciparum* malaria: differential adhesion of infected erythrocytes to CD36 and ICAM-1. *J. Infect. Dis.* **164**, 163–169 (1991).
62. Turner, G. D. *et al.* An immunohistochemical study of the pathology of fatal malaria. Evidence for widespread endothelial activation and a potential role for intercellular adhesion molecule-1 in cerebral sequestration. *Am. J. Pathol.* **145**, 1057–1069 (1994).
63. Pain, A. *et al.* A non-sense mutation in CD36 gene is associated with protection from severe malaria. *Lancet* **357**, 1502–1503 (2001).
64. Rowe, J. A., Moulds, J. M., Newbold, C. I. & Miller, L. H. *P. falciparum* rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1. *Nature* **388**, 292–295 (1997).
65. Heddini, A. *et al.* Fresh isolates from children with severe *Plasmodium falciparum* malaria bind to multiple receptors. *Infect. Immun.* **69**, 5848–5856 (2001).
66. Pain, A. *et al.* Platelet-mediated clumping of *Plasmodium falciparum*-infected erythrocytes is a common adhesive phenotype and is associated with severe malaria. *Proc. Natl Acad. Sci. USA* **98**, 1805–1810 (2001).
67. Fried, M., Nosten, F., Brockman, A., Brabin, B. J. & Duffy, P. E. Maternal antibodies block malaria. *Nature* **395**, 851–852 (1998).
68. Bull, P. C. *et al.* Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nature Med.* **4**, 358–360 (1998).
69. Bull, P. C., Lowe, B. S., Kortok, M. & Marsh, K. Antibody recognition of *Plasmodium falciparum* erythrocyte surface antigens in Kenya: evidence for rare and prevalent variants. *Infect. Immun.* **67**, 733–739 (1999).
70. Bull, P. C. *et al.* *Plasmodium falciparum*-infected erythrocytes: agglutination by diverse Kenyan plasma is associated with severe disease and young host age. *J. Infect. Dis.* **182**, 252–259 (2000).
71. Gupta, S., Snow, R. W., Donnelly, C. A., Marsh, K. & Newbold, C. Immunity to non-cerebral severe malaria is acquired after one or two infections. *Nature Med.* **5**, 340–343 (1999).
72. Clark, I. A. & Cowden, W. B. Why is the pathology of falciparum worse than that of vivax malaria? *Parasitol. Today* **15**, 458–461 (1999).
73. Wong, D. & Dorovini-Zis, K. Upregulation of intercellular adhesion molecule-1 (ICAM-1) expression in primary cultures of human brain microvessel endothelial cells by cytokines and lipopolysaccharide. *J. Neuroimmunol.* **39**, 11–21 (1992).
74. Levesque, M. C. *et al.* Nitric oxide synthase type 2 promoter polymorphisms, nitric oxide production, and disease severity in Tanzanian children with malaria. *J. Infect. Dis.* **180**, 1994–2002 (1999).
75. Dobbie, M., Crawley, J., Waruiru, C., Marsh, K. & Surtees, R. Cerebrospinal fluid studies in children with cerebral malaria: an excitotoxic mechanism? *Am. J. Trop. Med. Hyg.* **62**, 284–290 (2000).
76. Schofield, L. & Hackett, F. Signal transduction in host cells by a glycosylphosphatidylinositol toxin of malaria parasites. *J. Exp. Med.* **177**, 145–153 (1993).
77. Tachado, S. D. *et al.* Glycosylphosphatidylinositol toxin of *Plasmodium* induces nitric oxide synthase expression in macrophages and vascular endothelial cells by a protein tyrosine kinase-dependent and protein kinase C-dependent signaling pathway. *J. Immunol.* **156**, 1897–1907 (1996).
78. Naik, R. S. *et al.* Glycosylphosphatidylinositol anchors of *Plasmodium falciparum*: molecular characterization and naturally elicited antibody response that may provide immunity to malaria pathogenesis. *J. Exp. Med.* **192**, 1563–1576 (2000).
79. Toure-Balde, A. *et al.* *Plasmodium falciparum* induces apoptosis in human mononuclear cells. *Infect. Immun.* **64**, 744–750 (1996).
80. Hirunpetcharat, C. & Good, M. F. Deletion of *Plasmodium berghei*-specific CD4⁺ T cells adoptively transferred into recipient mice after challenge with homologous parasite. *Proc. Natl Acad. Sci. USA* **95**, 1715–1720 (1998).
81. Modiano, D. *et al.* Different response to *Plasmodium falciparum* malaria in West African sympatric ethnic groups. *Proc. Natl Acad. Sci. USA* **93**, 13206–13211 (1996).
82. Modiano, D. *et al.* Baseline immunity of the population and impact of insecticide-treated curtains on malaria infection. *Am. J. Trop. Med. Hyg.* **59**, 336–340 (1998).

Acknowledgements

We thank J. Barnwell for help with Fig. 2 and for sharing data before publication.

The economic and social burden of malaria

Jeffrey Sachs*† & Pia Malaney*

*Center for International Development, John F. Kennedy School of Government, Harvard University, 79 John F. Kennedy St., Cambridge, Massachusetts 02138, USA (e-mail: pia_malaney@harvard.edu)

†Commission on Macroeconomics and Health, World Health Organization, Avenue Appia 20, 1211 Geneva 27, Switzerland

Where malaria prospers most, human societies have prospered least. The global distribution of per-capita gross domestic product shows a striking correlation between malaria and poverty, and malaria-endemic countries also have lower rates of economic growth. There are multiple channels by which malaria impedes development, including effects on fertility, population growth, saving and investment, worker productivity, absenteeism, premature mortality and medical costs.

Long before economists attempted to estimate the costs of malaria, natural selection had already demonstrated the phenomenal burden of the disease. Certain genetic polymorphisms, such as sickle cell trait, were selected for because of their protective effect against malaria when inherited from one parent, even though the same allele inherited from both parents is fatal. In essence, the chance of death from malaria was so high as to justify welcoming a potentially fatal mutation into the gene pool^{1–3}.

Given this evolutionary backdrop, it would indeed be surprising if the economic and demographic toll of malaria were not comparably dramatic. Sadly, malaria does little to disappoint. The numbers are staggering: there are 300 to 500 million clinical cases every year, and between one and three million deaths, mostly of children, are attributable to this disease⁴. Every 40 seconds a child dies of malaria, resulting in a daily loss of more than 2,000 young lives worldwide. These estimates render malaria the pre-eminent tropical parasitic disease and one of the top three killers among communicable diseases.

Although the last century witnessed many successful programmes at country level to eliminate the parasite, the world is now facing a rapidly increasing disease burden⁵. This has been attributed to several causes, including population movements into malarious regions, changing agricultural practices including the building of dams and irrigation schemes, deforestation, the weakening of public health systems in some poor countries, and more speculatively, long-term climate changes such as more pronounced El Niño cycles and global warming. Furthermore, resistance to drugs and insecticides used to counter this disease has been evolving in tandem with growing caseloads. With a rapidly growing population in regions with high malaria transmission, it has been estimated that in the absence of effective intervention strategies the number of malaria cases will double over the next 20 years⁴.

Global transmission patterns

The malaria burden is not evenly distributed. The global pattern of malarial transmission suggests a disease centred in the tropics, but with a reach into subtropical regions in five continents. Attempts to eliminate or at least suppress the disease have been an important public health story

through much of the last century. At malaria's furthest reaches, in temperate zones characterized by strong seasonality and cold winters, these attempts have been successful. Beyond any other factors, this reflects the fact that the base case reproduction rate of malaria is considerably lower in temperate regions than in the tropics, so that moderately intensive efforts at vector control and case management can lead to elimination of the disease. The remarkably high transmission rates in sub-Saharan Africa also reflect the particular capacity of Africa's main vector mosquitoes, the *Anopheles gambiae* complex of species, with their remarkable tendency towards human biting (anthropophily).

These climatic patterns of course reflect the natural history of the disease. The malaria parasite is transmitted to the female *Anopheles* mosquito from an infected individual when it takes a blood meal as a prelude to the reproductive process. The parasite must undergo a life-cycle change within the mosquito before it becomes infectious to other individuals in the course of subsequent blood meals. The period required for that life-cycle change increases as the ambient temperature declines, and given the life span of the mosquito, transmission becomes much less likely when the temperature falls below 18 °C. Moreover, malaria parasites cease development completely at temperatures below 16 °C, and many species of vector mosquitoes suspend biting activity at very low temperatures, further reducing the stability of malaria transmission in temperate regions^{6–8}.

Although other climatic features such as rainfall and humidity also affect the stability of transmission⁸, seasonal temperature variation is a predominant factor in explaining the geographical distribution of the disease. Cold winters facilitated effective elimination of malaria infection from much of the temperate zone, leading malariologist Battista Grassi in 1901 to declare malaria a "giant with feet of clay", an obstacle that can readily be eliminated once appropriate interventions become available.

With the benefit of hindsight, we now understand that this optimistic statement is at best a statement concerning temperate-zone malaria. In tropical regions, exposure to mosquitoes may be perennial and frequently includes several contacts with infected vector mosquitoes each night. Such inoculation rates, combined with the long

duration of parasite survival in the host, rapidly saturate local human populations, resulting in universal prevalence and superinfection. This stable pattern of transmission resists amelioration, and vector control efforts that succeed in temperate zones have repeatedly failed to eradicate the parasite from tropical and subtropical regions, although control is possible (ref. 7; and see below). The changing global pattern of malaria transmission from 1946 to 1994 illustrates the success of antimalaria efforts in the more temperate regions of the world and the increased concentration of disease burden in the tropics (Fig. 1). Today, Africa alone accounts for 90% of malaria mortality.

The relationship between poverty and malaria

As a general rule of thumb, where malaria prospers most, human societies have prospered least. The global distribution of per-capita gross domestic product (GDP) in 1995, adjusted for purchasing power, shows a striking and unmistakable correlation between malaria and poverty (Fig. 2). Poverty is concentrated in the tropical and subtropical zones, the same geographical boundaries that most closely frame malaria transmission. The extent of the correlation suggests that malaria and poverty are intimately related. In fact, a comparison of income in malarious and non-malarious countries indicates that average GDP (adjusted to give parity of purchasing power) in malarious countries in 1995 was US\$1,526, compared with US\$8,268 in countries without intensive malaria — more than a fivefold difference⁹. Malaria-endemic countries are not only poorer than non-malarious countries, but they also have lower rates of economic growth. Between 1965 and 1990, countries in which a large proportion of the population lived in regions with *Plasmodium falciparum* malaria experienced an average growth in per-capita GDP of 0.4% per year, whereas average growth in other countries was 2.3% per year⁹.

This correlation can, of course, be explained in several possible ways. Poverty may promote malaria transmission; malaria may cause poverty by impeding economic growth; or causality may run in both directions. It is also possible that the correlation is at least partly spurious, with the tropical climate causing poverty for reasons unrelated to malaria. We tend to favour the explanation that causation runs in both directions, with the causal link from malaria to underdevelopment much more powerful than is generally appreciated.

It is certainly true that poverty itself can be held accountable for some of the intense malaria transmission recorded in the poorest countries. Personal expenditures on prevention methods such as bednets or insecticides, increased funding for government control

programmes, and general development such as increased urbanization can reduce malaria transmission. The elimination of malaria from wealthier countries in the 1930s to 1950s, such as the United States, Italy, Greece and Spain, was a result of both socioeconomic development and intensive antimalaria interventions. Improved housing, especially the provision of screened doors and windows, limits contact between mosquitoes and people. This, combined with efforts at environmental management such as the draining of swampland, which eliminates the breeding grounds of certain vector mosquitoes, and indoor spraying of residual insecticides such as DDT, successfully eliminated malaria from most temperate zone countries¹⁰.

But economic development alone is not enough. Even relatively wealthy countries with high year-round temperatures, such as Oman and the United Arab Emirates, have been unable to eliminate the disease. And within malarious regions, people from relatively wealthy households frequently fall ill from malaria. When malaria incidence within African villages is stratified by household income, there is often little difference across income classes¹¹.

The causation in the other direction, from malaria to poverty, also seems to be robust and powerful. Cross-country regression analysis estimating the long-term impacts of malaria on economic growth and development suggest the significance of the economic burden of the disease⁹. This analysis finds that countries in which a high proportion of the population lived in regions of *P. falciparum* malaria transmission in 1965 had annual economic growth rates that were 1.3% lower than other countries over the period 1965–1990, even after controlling for the other standard growth determinants used in macroeconomic analyses. These other determinants include levels of human capital, life expectancy, initial income, and macroeconomic policy indicators of various kinds as well as geographical factors such as tropical location that could be simultaneously influencing malaria and economic growth. Because this shortfall refers to the annual growth rate, the long-term effect on the level of gross national product (GNP) per capita is the cumulative effect of an annual reduction in growth. In the mathematical formulation used by Gallup and Sachs⁹, the long-term effect can be estimated as reducing the level of GNP per capita in a malarious country by more than half in comparison with a non-malarious country. Table 1 shows the estimated impact of malaria on incomes in thirty-one African countries between 1980 and 1995.

Malaria represents broad social and economic costs

Traditionally, studies that have attempted to estimate the economic burden of malaria have focused on the private and non-

Figure 1 Global distribution of malaria. The changing global distribution of malaria risk from 1946 to 1994 shows a disease burden that is increasingly being confined to tropical regions.

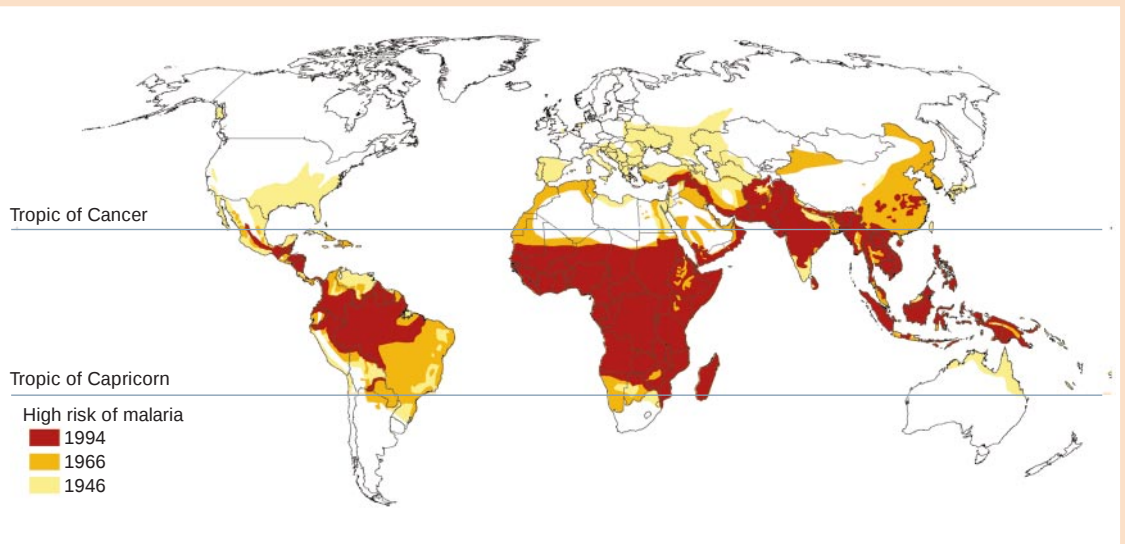
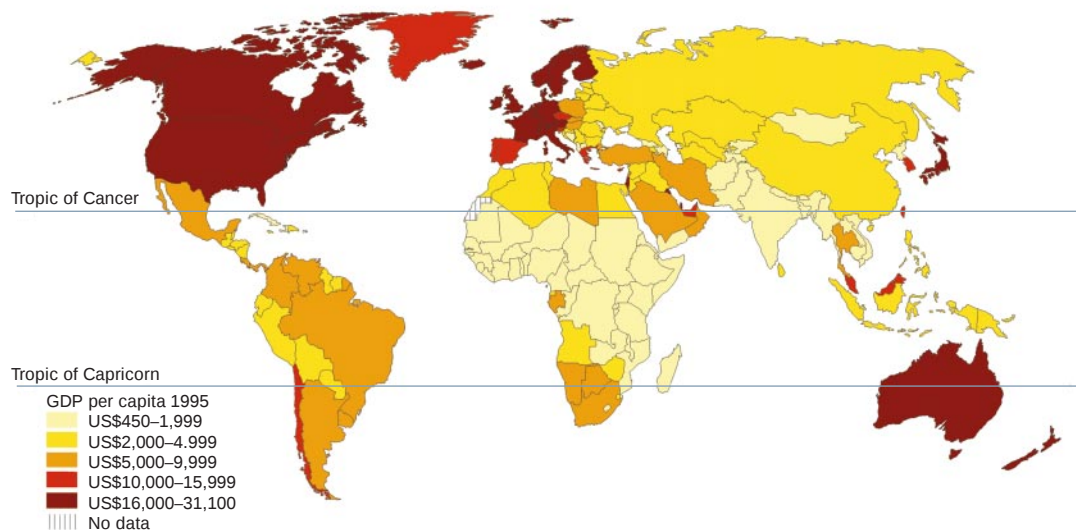


Figure 2 Global distribution of per capita GDP. The global pattern of income distribution is highly uneven, with average income levels significantly lower in tropical regions.



private medical costs associated with the disease, as well as some measure of the income that is foregone as a result of malaria morbidity and mortality. Private medical costs refer to personal expenditures on prevention, diagnosis, treatment and care of the disease. They include such factors as expenditure on bednets, doctor's fees, the cost of anti-malarial drugs, and the cost of transportation to medical facilities and the necessary support provided there. Non-private medical care costs are public expenditures on both prevention and treatment of the disease. They include expenditures by the government on such factors as vector control, health facilities, education and research. Foregone income is generally estimated by calculating the value of lost workdays as a result of malaria and malaria-related illness, based on estimated wages. In the case of mortality, foregone income is estimated by calculating the capitalized value of future lifetime earnings that would have been earned by those who died prematurely as a result of the disease, based on projected incomes for different age groups, basic longevity data and age-specific mortality rates.

These studies have found a burden that is significant and especially severe for those in the lowest income brackets. But the estimates, which average approximately 1% of GDP, simply miss some of the most important ways in which the disease affects long-term economic growth and development. In effect, traditional studies have used accounting techniques which assume that the economic costs of malaria can be determined by the average cost of an individual episode of illness, multiplied by the total number of cases encountered, and adding any fixed costs expended in prevention and treatment. Such techniques might be appropriate when there are a few episodes of disease (for example, episodes of malaria in the United States and Europe resulting from travel in malarious regions), but make little sense when extended to situations of high transmission.

There are at least two broad categories of mechanisms through which malaria can impose economic costs well beyond direct medical costs and foregone incomes. The first is the effects that occur through changes in household behaviour in response to the disease, which can result in broad social costs. These include such factors as schooling, demography, migration and saving. The second are macroeconomic costs that arise specifically in response to the pandemic nature of the disease and that cannot be assessed at a household level. These include the impact of malaria on trade, tourism and foreign direct investment. Below we

explore some of these additional pathways through which malaria affects economic development.

Long-term demographic consequences

Malaria kills more than one million people a year, and perhaps close to three million when the role of malaria in deaths related to other diseases is included. Much of the mortality in endemic areas is concentrated among children under the age of five. In areas of stable endemic transmission about 25% of all-cause mortality in children aged 0 to 4 years has been attributed directly to malaria¹². Furthermore, trials using insecticide-treated bednets in some African countries have shown a reduction in all-cause mortality among infants and children of up to 60%, indicating that in these areas malaria directly and indirectly accounts for an extremely high proportion of infant and child deaths^{13–16}.

Aside from the direct demographic consequences of this alarming level of mortality there are also likely to be significant indirect effects. Historical evidence has shown that high infant and child mortality rates are linked closely to high fertility rates. Along with other factors such as household income, female education and the availability of birth control, infant and child mortality are important factors in the fertility decisions of households^{17–20}. The simplest explanation for this link is that parents have additional children to replace the ones that they lose. Another hypothesis, known as the 'child-survivor hypothesis', is that parents base their fertility decisions on a desire for a certain number of surviving children (for example, to guarantee at least one surviving male heir, or one surviving child into the old age of the parents). In this theory, risk-averse households raise fertility by even more than expected mortality, in order to ensure a sufficiently high likelihood of the desired number of surviving children. This theory predicts that a high burden of malaria will lead to a disproportionately high fertility rate and an overall high population growth rate in regions of intense malaria transmission. These predictions are supported by cross-country evidence, although the direct causal linkages from malaria deaths to increased fertility to rapid population growth is circumstantial, and yet to be proved.

A high fertility rate among poor households is also likely to lead to reduced investments in education per child, a phenomenon that economists term the 'quantity–quality trade-off'. A high-fertility environment can also entail especially large human capital costs for women. When women have very high fertility rates, parents may choose to invest less in the education of their daughters, knowing

that they are likely to spend a considerable portion of their working years involved in child-rearing activities rather than in the labour force where they would reap the economic returns to education. Moreover, while the hours they spend caring for a young child may not represent a majority of their workday, women will undoubtedly be limited in their employment choices by the need to be available to the child, with a resulting loss in work opportunities and job experience. Over the long term, such factors can have a sizeable impact on economic growth and productivity.

While it is difficult to estimate the exact responsiveness of fertility and education to malaria-induced mortality, a high-fertility/high-mortality environment presents very high costs at both a household and a national level. At the simplest level, the economic value of resources invested in infants who do not survive can be significant. Estimates of the number of hours parents spend in child-rearing for every year of a child's life and calculations of the productive time 'lost' based on infant and child mortality in high-mortality societies show this to be a substantial cost²¹.

Additional effects will operate through changes in the dependency ratio — the number of dependants per working-age population. If couples are following a child-survivor strategy, at any given time there will be many more young children in the population than are expected to survive into adulthood. This directly lowers GNP per capita (since GNP is produced by adults, whereas GNP per capita is measured relative to the total population). A high dependency ratio has also been shown to have long-term growth impacts through other channels, such as household saving behaviour²².

The acquisition of human capital

Where malaria is highly endemic, adults generally develop partial immunity to the symptoms of the disease. Young children, however, bear a considerable burden in terms of malaria morbidity and mortality. Although this morbidity is most concentrated among pre-school children, school-age children also suffer the effects, resulting in school absenteeism. For example, in Kenya it was found that primary school students miss 11% of school days per year because of malaria, and secondary school students miss 4.3% of school days²³. Another study attributed 13–50% of medically related school absences to the disease²⁴. The adverse effects on schooling are likely to go far beyond the number of days lost per year, as absenteeism increases failure rates, repetition of school years, and drop-out rates.

An even more severe consequence can arise from the impact of malaria on cognitive development and learning ability. Although there is some debate about the direct relationship between malaria and mental functioning, a number of channels have been identified through which malaria can affect cognitive abilities in ways ranging from subtle to profound²⁵. For example, children with malaria are found to have poorer nutritional status than non-malarial children^{26,27}, an outcome that can impair brain development^{28–30}. In at least one area of cognitive development, the performance of fine motor functions, it has been found that parasitaemic children perform worse than non-parasitaemic children³¹.

At an even earlier stage, malaria impacts on fetal development. Pregnant women are particularly likely to be infected by malaria as a result of diminished immunity, and malaria-related anaemia of the mother is related to low birthweight among babies. This is a risk factor for neurosensory, cognitive and behavioural development of children³². Compared with normal birthweight babies, low birthweight babies are two to four times more likely to experience failure in school³³.

There can also be long-term cognitive effects of severe cases of malaria. Cerebral malaria affects approximately 575,000 children a year in Africa, killing 10–40% of patients^{25,34,35}. Of those who survive, 5–20% experience neurological sequelae including behavioural disorders and impairment in the ability to carry out

Table 1 Loss from the economic growth penalty of malaria endemicity in 31 African countries 1980–1995

Country	Aggregate loss (PPP-adjusted US\$ million)*	Per person loss (PPP-adjusted US\$)*	Fraction of actual 1995 income
Benin	1,172	214	18%
Botswana	503	347	5%
Burkina Faso	1,684	162	18%
Burundi	730	117	18%
Cameroon	4,227	318	18%
Central African Rep.	884	270	18%
Chad	995	154	17%
Congo	759	288	18%
Congo, Dem. Rep.	7,125	162	18%
Côte d'Ivoire	4,107	294	18%
Gabon	1,389	1,290	17%
Gambia	251	226	18%
Ghana	5,355	314	18%
Guinea Bissau	152	142	14%
Kenya	5,272	198	18%
Lesotho	0	0	0%
Madagascar	2,280	167	18%
Malawi	1,072	110	18%
Mali	1,222	125	17%
Mauritania	611	269	15%
Mauritius	0	0	0%
Namibia	832	539	10%
Niger	1,457	161	17%
Nigeria	17,315	156	18%
Rwanda	656	102	18%
Senegal	2,426	286	18%
Sierra Leone	366	87	17%
South Africa	4,056	98	1%
Togo	1,166	285	18%
Zambia	1,359	151	18%
Zimbabwe	4,214	383	18%
Total	73,638	185	10%

*Figures are reported in US dollars held constant at 1987 prices and are adjusted for purchasing power parity (PPP), that is, for differences in the local purchasing power of the currency with respect to a fixed market basket of goods and services.

executive functions such as initiating, planning and executing tasks^{25,36,37}. Furthermore, seizures that result from clinical malaria are associated with increased risk of impaired intellectual functioning²⁵.

Human capital constitutes a key factor in economic growth. Modern economic theories have posited that returns to human capital are in fact increasing in scale, with the total impact on economic growth being greater than the sum of the individual contributions. The impact of malaria on economic growth rates through the mechanism of depressing the rate of human capital accumulation could be considerable. Although there have been some estimations of the loss of educational investment expenditures as a result of lost school days³⁸, the overall impact of malaria on human capital development in children remains largely unexplored and unquantified.

The acquisition of physical capital

The direct costs of prevention and treatment of the disease eat into the disposable income of poor families, as do the costs of lost productivity. Whereas economic models suggest that increased risk of illness could, in fact, increase savings if families were trying to protect themselves from vulnerability to economic shocks by building a buffer, there is little evidence to support this idea in the poor, rural households that have been studied. The evidence suggests that malaria decreases household savings as families are forced to hire labour to compensate for days lost to morbidity³⁹.

The long-term demographic impacts of malaria also potentially affect savings rates. If reduced malaria-related infant and child mortality does bring down fertility rates, then the resulting decline in the dependency ratio could lead to a higher savings rate. It has been argued that families with lower dependency ratios, with fewer children to feed and clothe, are more capable of saving and investing, not just in physical capital or capital markets, but also in the human capital of each child through education, as noted above.

Malaria affects the movement of people

Risk of malaria infection has a profound effect on the mobility of human populations and the construction of new settlements, with consequent impact on economic growth and development⁴⁰. Although residents of highly endemic sites generally develop disease-modifying immunity that diminishes malaria-related morbidity and mortality, antigenic diversity may limit this effect geographically. Such acquired immunity, moreover, can be transient and is often lost within a year or so in the absence of re-infection, such as during a period of schooling or employment away from the malarious region. As in the case of migrants, the return of residents to their original endemic homes carries an increased risk of death or disease.

The importance of acquired adult immunity in protecting against malaria morbidity and mortality potentially inhibits the movement of labour between malarious and non-malarious regions. This represents an economic cost because human mobility permits labour to move to regions where it is most productive. By limiting such movement, malaria would interfere with skill matching and generally inhibit maximization of worker productivity. Furthermore, incentives to expand markets into malarious regions of the world will be lost in the event that trade and commercialization expose people to an increased burden of malaria, a factor that can hinder long-term economic development.

Trade and foreign direct investment

Throughout history, malaria has suppressed the economic linkages between malarious and non-malarious regions. In previous centuries this isolating effect may have served to protect malarious countries from European colonizers. But in today's globalized economy, the isolation has more negative effects⁴¹. Although direct measurement of the impact of disease on decision-making by foreign investors is difficult to gauge, failure of investment is likely to be one of the most costly effects of malaria with respect to long-term growth. Investors from non-malarious regions tend to shun malarious regions for fear of contracting the disease — a fear that is sadly well grounded in reality, as evidenced recently by the experience of Billiton, a London-based mining and metals company. In a US\$1.4 billion joint venture investment to build an aluminium smelter in Mozambique, the largest foreign investment so far in that country, the company was faced with 7,000 cases of malaria in two years, and the death of 13 expatriate employees⁴².

Industries such as tourism are particularly hard hit by malaria transmission, as is becoming clearer in countries such as Mozambique and South Africa as they attempt to encourage investment in these areas with limited success. Investments in all sorts of production — in mining, agriculture and manufacturing — may similarly be crippled if the labour force faces a heavy disease burden, or if the burden raises the costs of attracting the needed labour to a malarious region. In an economic era in which international trade and finance is critical for economic development, these adverse effects on foreign trade and investment are likely to be of tremendous macroeconomic importance.

When malaria transmission was suppressed in the subtropical regions of southern Europe in the 1950s, there was a subsequent surge of economic growth. Greece, Portugal and Spain all began a process of accelerated economic development. One of the sources of that growth was increased foreign investment by northern

European firms. Another was greatly increased tourism. Although the linkages from malaria control to increased investment and tourism have not been proven rigorously, the timing and sectoral evidence support the proposition that malaria control played an economically important role.

Malaria and other illnesses

One of the surprising results to emerge from large-scale trials of insecticide-treated bednets is that the reduction in all-cause mortality with the use of bednets is considerably greater than the reduction in malaria-attributed mortality. This might imply that malaria is closely linked to other diseases, either as a direct causal factor, or because malaria renders individuals more susceptible to other infections¹².

The indirect effects of malaria can begin even before birth. As mentioned earlier, pregnant women are at a higher risk of malaria infection, and malarial pregnancy can result in miscarriages, neonatal and infant mortality, low birthweight and congenital infection. Acute and chronic malaria infections can alter the immune system and the body's response to vaccines, and increase vulnerability to other infections.

Furthermore, chronic malaria is an important causal factor in anaemia^{27,43}, which has been shown to have direct physical effects, lowering worker productivity and output^{44,45}. Malaria is also associated with hyper-reactive malarial splenomegaly, chronic renal damage and the nephrotic syndrome, and Burkitt's lymphoma. Increasingly, malaria is becoming a factor in the transmission of human immunodeficiency virus (HIV), the virus that causes AIDS, as children with severe malaria often require blood transfusions, and much of the blood supply in sub-Saharan African countries is infected with HIV. To the extent that malaria is a factor in other illnesses, any assessment of the economic burden of malaria should include the range of costs that are associated with these illnesses.

Looking to the future

Although the failure of earlier attempts at global eradication of malaria led initially to donor fatigue and a weakening of intervention efforts, the lessons of history can serve us well as we look to the future in our battle against the disease. Experience indicates that temperate and tropical malarias can be considered as distinct epidemiological manifestations of a common array of parasites, and although elimination may be a reasonable goal in temperate zones, it has proved unrealistic in many tropical and subtropical environments⁴⁶. However, a range of cost-effective approaches is available to reduce the burden of malaria, including case management, the use of insecticide-treated bednets, indoor residual spraying, and environmental control measures such as larvaciding (controlling mosquitoes at the larval stage through the use of chemicals) and filling and draining of breeding sites⁴⁷. Each of these interventions seems to have great value with respect to the health gains achieved per dollar spent and, by extension, would have enormous economic benefits. Effective malaria control programmes can be developed using a combination of these approaches adapted to local needs based on specific ecological, epidemiological, economic and social conditions.

Within its dominion, malaria affects almost every aspect of social and economic endeavour, including fertility, savings and investment rates, crop choices, schooling and migration decisions. Where transmission is intense, the disease creates a complex set of biological and behavioural responses with a long-term effect on economic growth and development that goes well beyond the additive costs of individual cases. Suppressing malaria in poor, highly malarious regions, especially in sub-Saharan Africa, offers the potential to initiate a virtuous cycle in which improved health spurs economic growth, and rising income further benefits human health. The economic evidence also suggests that high priority targets should include port cities, potential tourist destinations,

mining operations and high-value-added agricultural settings. Even when disease incidence is lower in such settings than in other places, the economic benefits of disease control in these locations (through increased tourism, trade and investment) could be enormous. More generally, reducing malaria transmission, and other communicable diseases, in low-income settings may well prove to be among the most effective available spur to overall economic development.

Despite the enormous potential benefits of antimalaria programmes, the level of international spending on malaria control in poor regions has been dismal. While current spending has been running at less than US\$100 million per year, actual needs are more than an order of magnitude greater. Detailed costing estimates prepared by the World Health Organization (WHO) Commission on Macroeconomics and Health find that effective malaria prevention and treatment will require an additional US\$2.5 billion per year as of 2007, increasing to US\$4 billion per year as of 2015. In order to scale up the international effort on malaria control, the WHO and partner institutions have launched a campaign to 'Roll Back Malaria'. This important venture will probably be hampered in its mission as it is still considerably underfunded relative to the need for malaria control and the expected return on such investments in poorer countries. The newly announced Global Fund to Fight HIV/AIDS, Tuberculosis and Malaria may finally deliver some succour, but even that fund remains grossly underfinanced at this point. Nevertheless, we should be optimistic that the new visibility for infectious disease symbolized by such programmes represents but the first step towards putting appropriate resources at the service of those fighting the global burden of this disease. □

1. Luria, S. E. 36 *Lectures in Biology* p. 439 (The MIT Press, Cambridge, 1975).
2. Luzzatto, L. Genetics of red cells and susceptibility to malaria. *Blood* **54**, 961–976 (1979).
3. Hill, A. V. S. *et al.* Molecular analysis of the association of HLA-B53 and resistance to severe malaria. *Nature* **360**, 434–439 (1992).
4. Breman, J. The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. *Am. J. Trop. Med. Hyg.* **64**(1,2), 1–11 (2001).
5. World Health Organization. Factsheet No. 94 (World Health Organization, Geneva, 1998).
6. Anderson, R. M. & May, R. *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, Tokyo, 1991).
7. Colluzzi, M. The clay feet of the malaria giant and its African roots: hypotheses and inferences about origin, spread and control of *Plasmodium falciparum*. *Parasitologia* **41**, 277–283 (1999).
8. Gilles, H. M. & Warrell, D. A. *Bruce-Chwatts Essential Malariology* (Arnold, Boston, 1993).
9. Gallup, J. & Sachs, J. The economic burden of malaria. *Am. J. Trop. Med. Hyg.* **64**(1,2), 85–96 (2001).
10. Kitron, U. & Spielman, A. Suppression of transmission of malaria through source reduction: antianopheline measures applied in Israel, the United States and Italy. *Rev. Infec. Dis.* **11**, 391–406 (1989).
11. Filmer, D. Fever and its treatment in the more and less poor in sub-Saharan Africa. Development Research Group (The World Bank, Washington DC, 2000).
12. Snow, R. W., Craig, M., Deichmann, U. & Marsh, K. Estimating mortality, morbidity and disability due to malaria among Africa's non-pregnant population. *Bull. World Health Org.* **77**, 624–640 (1999).
13. Alonso, P. L. *et al.* The effect of insecticide-treated bed nets on mortality of Gambian children. *Lancet* **337**, 1499–1502 (1991).
14. D'Alessandro, U. *et al.* Mortality and morbidity from malaria in Gambian children after introduction of an impregnated bednet programme. *Lancet* **345**, 479–483 (1995).
15. Nevill, C. G. *et al.* Insecticide-treated bednets reduce mortality and severe morbidity from malaria among children on the Kenyan coast. *Trop. Med. Int. Health* **1**, 139–146 (1996).
16. Binka, F. N. Impact of permethrin impregnated bednets on child mortality in Kassena-Nankana district, Ghana: a randomized controlled trial. *Trop. Med. Int. Health* **1**, 147–154 (1996).
17. Yamada, T. Causal relationships between infant mortality and fertility in developed and less developed countries. *South. Econ. J.* **52**, 364–371 (1985).
18. Galloway, P. R., Lee, R. D. & Hammel, E. A. in *From Death to Birth: Mortality Decline and Reproductive Change* (eds Montgomery, M. R. & Cohen, B.) 182–226 (National Academy Press, Washington DC, 1998).

19. Handa, S. The impact of education, income, and mortality on fertility in Jamaica. *World Dev.* **28**, 173–186 (2000).
20. Rosensweig, M. & Schultz, T. P. Consumer demand and household production: the relationship between fertility and child mortality. *Am. Econ. Rev.* **73**, 38–43 (1983).
21. Reher, D. Wasted investments: some economic implications of childhood mortality patterns. *Popul. Stud.* **49**, 519–536 (1995).
22. Bloom, D. E., Canning, D. & Malaney, P. N. Demographic change and economic growth in Asia. *Popul. Dev. Rev.* **26**(Suppl.), 257–290 (2000).
23. Leighton, C. & Foster, R. Economic impacts of malaria in Kenya and Nigeria. Major Applied Research Paper no 6, HFS project (Abt Associates, Bethesda, 1993).
24. Brooker, S. *et al.* Situation analysis of malaria in school-aged children in Kenya: what can be done? *Parasitol. Today* **16**, 183–186 (2000).
25. Holding, P. A. & Snow, R. W. Impact of *Plasmodium falciparum* malaria on performance and learning: review of the evidence. *Am. J. Trop. Med. Hyg.* **64**(1,2), 68–75 (2001).
26. Rowland, M. G., Cole, T. J. & Whitehead, R. G. A quantitative study into the role of infection in determining nutritional status in Gambian village children. *Br. J. Nutr.* **37**, 441–450 (1977).
27. Shiff, C. *et al.* Changes in weight gain and anaemia attributable to malaria in Tanzanian children living under holoendemic conditions. *Trans. R. Soc. Trop. Med. Hyg.* **90**, 262–265 (1996).
28. Lozoff, B. Nutrition and behaviour. *Am. Psychiatry* **44**, 231–236 (1989).
29. McKay, H., Sinisterra, L., McKay, A., Gomez, H. & Lloreda, P. Improving cognitive ability in chronically deprived children. *Science* **200**, 270–278 (1978).
30. Grantham-McGregor, S. M., Powell, C. A., Walker, S. P. & Himes, J. H. Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican study. *Lancet* **338**, 1–5 (1991).
31. Al Serouri, A. W., Grantham-McGregor, S. M., Greenwood, B. & Costello, A. Impact of asymptomatic malaria parasitaemia on cognitive function and school achievement of schoolchildren in the Yemen Republic. *Parasitology* **121**, 337–345 (2000).
32. McCormick, M. C., Brooks-Gunn, J., Workman-Daniels, K., Turner, J. & Peckmah, G. J., The health and development status of very low-birth-weight children at school-age. *J. Am. Med. Assoc.* **267**, 2204–2208 (1992).
33. Taylor, H. G. in *Early Brain Damage: Research Orientations and Clinical Observation* (eds Almi, C. & Fingers, S.) 325–345 (Academic, New York, 1984).
34. Murphy, S. & Breman, J. Gaps in the childhood malaria burden in Africa: cerebral malaria, neurological sequelae, anaemia, respiratory distress, hypoglycemia, and complications of pregnancy. *Am. J. Trop. Med. Hyg.* **64**(1,2), 57–67 (2001).
35. Greenwood, B. *et al.* Mortality and morbidity from malaria among children in a rural area of The Gambia, West Africa. *Trans. R. Soc. Trop. Med. Hyg.* **81**, 478–486 (1987).
36. Brewster, D. R., Kwiatkowski, D. & White, N. J. Neurological sequelae of cerebral malaria in children. *Lancet* **336**, 1039–1043 (1990).
37. Holding, P. A., Stevenson, J., Pershu, N. & Marsh, K. Cognitive sequelae of malaria with impaired consciousness. *Trans. R. Soc. Trop. Med. Hyg.* **93**, 529–534 (1999).
38. Kere, N. K., Keni, J., Kere, J. F., Bobogare, A. & Webber, R. H. The economic impact of plasmodium falciparum malaria on education investment: a Pacific Island case study. *Southeast Asian J. Trop. Med. Public Health* **24**, 659–663 (1993).
39. Nur, E. The impact of malaria on labour use and efficiency in the Sudan. *Soc. Sci. Med.* **37**, 1115–1119 (1993).
40. Sawyer, D. Economic and social consequences of malaria in new colonization projects in Brazil. *Soc. Sci. Med.* **37**, 1131–1136 (1993).
41. Spielman, A. & DiAntonio, M. *Mosquito* p. 247 (Hyperion, New York, 2001).
42. Choice of evils: as a tropical surge makes a comeback, so, too, does DDT. *Wall St. J.* 26 July (2001).
43. Hedberg, K. *et al.* *Plasmodium falciparum*-associated anaemia in children at a large urban hospital in Zaire. *Am. J. Trop. Med. Hyg.* **48**, 365–371 (1993).
44. Scholz, B. D., Gross, R., Schultink, W. & Sastroamidjojo, S. Anaemia is associated with reduced productivity of women workers even in less-physically-strenuous tasks. *Br. J. Nutr.* **77**, 47–57 (1997).
45. Basta, S., Soekirman, S., Karyadi, D. & Scrimshaw, N. S. Iron deficiency anaemia and the productivity of adult males in Indonesia. *Am. J. Clin. Nutr.* **32**, 916–925 (1979).
46. Spielman, A., Kitron, U. & Pollack, R. Time limitation and the role of research in the worldwide attempt to eradicate malaria. *J. Med. Entomol.* **30**, 6–19 (1993).
47. Goodman, C. A., Coleman, P. G. & Mills, A. J. Cost-effectiveness of malaria control in sub-Saharan Africa. *Lancet* **354**, 378–385 (1999).

Acknowledgements

This work was carried out under Program on Malaria, Economics and Human Affairs, directed by A. Spielman at the Center for International Development. This article has benefited greatly from Dr Spielman's support and ongoing tutelage in the scientific and epidemiological aspects of the malaria disease and vector. We are also grateful to E. Weinstein for valuable insights, to J. Gallup for previous collaboration, and to A. Kiszewski for sharing his entomological expertise. Excellent research assistance was provided by E. Gummerson.

Medical need, scientific opportunity and the drive for antimalarial drugs

Robert G. Ridley

Medicines for Malaria Venture, Box 1826, CH-1215 Geneva 15, Switzerland (e-mail: ridleyr@mmv.org)

Continued and sustainable improvements in antimalarial medicines through focused research and development are essential for the world's future ability to treat and control malaria. Unfortunately, malaria is a disease of poverty, and despite a wealth of scientific knowledge there is insufficient market incentive to generate the competitive, business-driven industrial antimalarial drug research and development that is normally needed to deliver new products. Mechanisms of partnering with industry have been established to overcome this obstacle and to open up and build on scientific opportunities for improved chemotherapy in the future.

When contemplating any review of chemotherapy it is important to recognize that there is no such thing as a perfect drug. All pharmaceutical agents have their liabilities as well as their benefits. The adverse effects of some drugs, such as thalidomide¹, have had a big impact on the types of study that are now required to obtain regulatory approval to market drugs. Regulatory studies place great emphasis on drug safety and on quality assurance of product, as well as on efficacy. 'Pharmacovigilance' — the continued monitoring and surveillance of drug safety after marketing authorization has been granted — is also essential to good pharmaceutical practice.

For infectious diseases, an additional liability and a chief motivation for continued innovation is the development of drug resistance. Malaria had been sheltered for many years from the dangers of resistance because of the outstanding properties of chloroquine and the slow speed at which resistance developed to this drug². But the final arrival of resistance to chloroquine on a global scale has exposed the ease with which resistance may develop to other drugs such as the antifolates^{3,4}. Figure 1 illustrates the dilemma that now faces malaria control authorities,

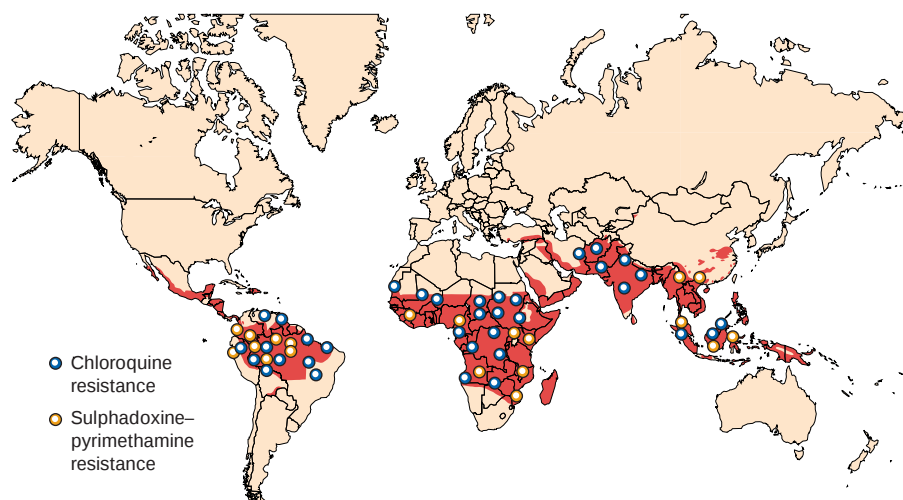
with global resistance prevailing against the two most widely used antimalarial drugs, chloroquine and the antifolate sulphadoxine/pyrimethamine. The challenge ahead lies in determining the best alternative therapies available for use now, the best prospects for drug development, regulatory approval and use in the short term, and the establishment of mechanisms and projects to ensure that improved drugs are sustainably discovered and developed into the future.

For brevity, I review primarily work carried out on the most significant of the four species of malaria parasite that infect humans, *Plasmodium falciparum*, and do not cover work on *Plasmodium vivax*, *Plasmodium ovale* or *Plasmodium malariae* in detail. I also focus on drugs and research directed against the erythrocytic stage of the parasite life cycle — the stage that gives rise to the clinical symptoms of the disease. I do not address the liver stage of infection.

Benefits and liabilities of existing antimalarial drugs

A good starting point for anybody interested in the detailed clinical properties of existing antimalarial drugs and their impact on antimalarial drug policy is a series of reports by the WHO (World Health Organization) on the use of antimalarial drugs⁵⁻⁷. The molecular aspects of malaria

Figure 1 Global status of resistance to chloroquine and sulphadoxine/pyrimethamine, the two most widely used antimalarial drugs. Data are from the WHO.

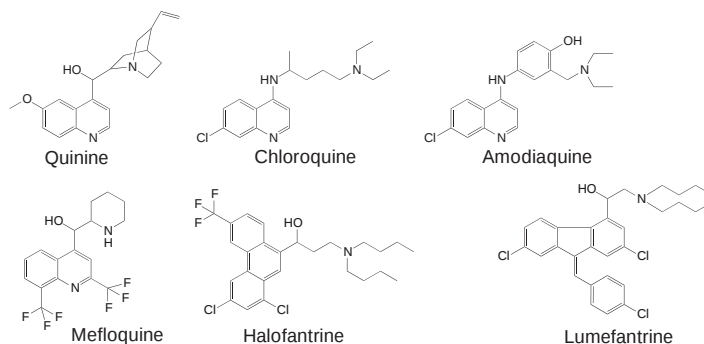


Box 1

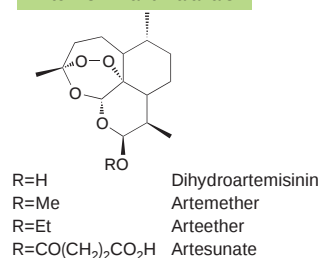
Overview of antimalarial drugs

Drug	Main limitations
Chloroquine Quinine Amodiaquine Mefloquine Halofantrine	Resistance Compliance/safety/resistance Safety/resistance (Safety)/resistance/(cost) Safety/resistance/cost
Artemisinins (artemether, arteether, artesunate)	Compliance/(safety)/(GMP)/ (cost)
Sulphadoxine– pyrimethamine	Resistance
Atovaquone– proguanil	Resistance potential/cost
Lumefantrine– artemether	(Compliance)/resistance potential/(cost)

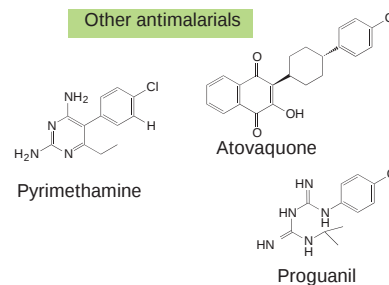
Quinoline and related antimalarials



Artemisinin antimalarials



Other antimalarials



Drugs that are currently in use as antimalarials have many attributes, but also possess certain liabilities that might be improved by further drug discovery and development. Liabilities placed in brackets refer to issues that are less serious for the drugs in question than those liabilities not placed in brackets. The structures of the main antimalarial drugs are provided for reference.

chemotherapy are well covered by a collection of articles in ref. 8. The principal antimalarial drugs in use are listed in Box 1.

Quinolines

Quinoline antimalarials and related aryl alcohols owe their origins to quinine — an active ingredient of *Cinchona* bark, which was first imported into Europe from Peru for antimalarial use in the seventeenth century⁹. Quinine has liabilities associated with toxicity (such as tinnitus) and, because it requires thrice daily administration over 7 days, this can result in poor compliance. The dependence on raw material for its extraction and the opportunities presented by its structural elucidation led to the development of the fully synthetic 4-aminoquinoline antimalarials — notably chloroquine and later amodiaquine^{10–12}, which are inexpensive and administered over 3 days.

Amodiaquine use has been limited since the mid 1980s after it was linked causally to the occurrence of occasional agranulocytosis in adult travellers taking the drug prophylactically⁵. But because amodiaquine retains a high degree of efficacy against all but the most highly chloroquine-resistant strains, there has been a recent increase in its use⁵. Mefloquine and halofantrine are structurally related drugs that are active against chloroquine-resistant strains, but resistance can develop rapidly to each of these drugs^{2,13}. Additional limitations include occasional neuropsychiatric disturbances for mefloquine and the fact that halofantrine is strongly contra-indicated in people with a history of heart disease⁵.

Artemisinins

Several semisynthetic derivatives of artemisinin — the active ingredient of the Chinese herb 'qinghao' (*Artemisia annua*), which was used traditionally for treating fevers — have been used increasingly over the past two decades^{5,7,14–17}. These derivatives (Box 1) include

artemether, arteether and artesunate, which are all metabolized to dihydroartemisinin — the main active agent in the body. These drugs are fast acting and act against gametocytes, the sexual stages of the parasite that infect mosquitoes.

The short half-lives of both the parent semisynthetic derivatives and the dihydroartemisinin metabolite necessitate treatment over 5–7 days when these compounds are used alone. They are therefore being used increasingly in combination with longer half-life drugs to reduce treatment time and increase individual compliance⁶. It is anticipated that rapid clearance of the parasites by artemisinin derivatives¹⁸ will reduce the chances of resistance development to the partner drugs¹⁹. The combination of artesunate with mefloquine is recommended policy in some parts of Thailand^{5,20}, and a study is nearing completion that involves numerous artesunate combinations in an attempt to identify the appropriate combinations for different epidemiological situations^{5,6}.

The universal applicability and practicability of this approach still needs to be assessed fully, however, especially in Africa^{6,21} where three main issues arise. First, what drug should be used for the longer half-life component of the combination? A partner drug with a very long half-life may be unsuitable for extended use in areas of intense transmission, where subsequent infections can occur before the drug has been cleared from the circulation but after the artemisinin derivative has been cleared; resistance is more likely to develop against the partner drug in such a situation. Second, compliance to complex drug regimens must be ensured against a backdrop of poor health-care infrastructure and patient supervision. The development of fixed-dose combinations (two or more compounds in a single tablet) might alleviate this problem. Third, the cost should be minimal. Although inexpensive by international standards, artemisinin derivatives are significantly more expensive than traditional antimalarials such as chloroquine and sulphadoxine/pyrimethamine.

The use of artemisinin derivatives has been negatively impacted by the observation that high parenteral doses of certain compounds can produce a limited, unique, selective brain stem neuronopathy in laboratory animals²². An analysis has concluded that these observations do not justify non-use of the compounds²³, and no associated clinical symptoms have been detected²⁴; however, limited preclinical data and quality assurance specifications for some products on the market continue to fuel discussion. Occasional allergic reactions to artesunate, the most widely used artemisinin derivative, have also been reported²⁵.

The artemisinin fixed-dose combination product of lumefantrine/artemether, which is now registered widely, is among the most recently approved antimalarial drugs. The aryl alcohol lumefantrine is similar to mefloquine and halofantrine, and no neurotoxicity was seen during animal toxicology studies in preclinical development²⁶. It is unclear whether a 2-day (4-dose) regimen or a 3-day (6-dose) regimen is the most appropriate treatment in Africa⁶. The need for higher dosing to ensure efficacy is supported by evidence of the cross-resistance of lumefantrine with mefloquine in both Thailand²⁷ and Cameroon²⁸, and the dependence on food for optimal adsorption of lumefantrine from the gut²⁹. With respect to the three issues on artemisinin combinations outlined above, it will be highly informative to discover how this drug performs in an African context.

Antifolates

The antifolate class of antimalarial drugs is not derived from plants and owes its origins to compounds generated through a knowledge of cell biology and synthetic medicinal chemistry. Fully reduced folate cofactors are essential for the key one-carbon transfer reactions needed for nucleotide biosynthesis and amino-acid metabolism³⁰. At present, the most significant antifolate used to treat malaria is undoubtedly the combination of the 2,4-diaminopyrimidine pyrimethamine, an inhibitor of dihydrofolate reductase (DHFR), and sulphadoxine, a sulphonamide that interferes with the action of dihydropteroate synthase (DHPS), another enzyme in the folate pathway.

The two components of the sulphadoxine/pyrimethamine combination act as synergists with each other, enhancing their activity and reducing the propensity for resistance development. The long half-life of the components may also account for the fact that sulphadoxine/pyrimethamine is extremely useful for intermittent treatment in pregnancy³¹, although this link still remains to be proved. Unfortunately, resistance seems to develop rapidly when this combination is used extensively^{32,33}. An additional concern that prevents the use of sulphadoxine/pyrimethamine as a prophylactic agent is that occasional hypersensitivity to the sulphur component may give rise to Stevens–Johnson syndrome — a toxic epidermal necrolysis that results in painful blistering of the skin⁵.

Atovaquone/proguanil

Atovaquone/proguanil is a fixed-dose combination whose development³⁴ shows similarity to that of sulphadoxine/pyrimethamine. The hydroxynaphthoquinone atovaquone interferes with mitochondrial electron transport. Rapid resistance develops against atovaquone³⁴ owing to a point mutation in cytochrome *c* reductase³⁵. The addition of proguanil to atovaquone results in a synergistic activity that prevents the rapid development of resistance.

At present atovaquone/proguanil is used primarily as a prophylactic agent, and its price is too high for it to achieve widespread use in developing countries. One reason for this cost is the complexity of the synthetic route to atovaquone.

Antibiotics

Common antibiotics acting against bacterial protein synthesis such as tetracycline, doxycycline and clindamycin inhibit parasite growth and are being used increasingly in combination with other

antimalarial treatments to augment their activity⁵. In parts of south-east Asia, quinine plus tetracycline and quinine plus doxycycline are commonly used combinations. But their use in Africa is limited because both antibiotics are contra-indicated in children under 8 years of age⁵. Clindamycin is recommended in combination with other antimalarials in some situations.

These antibiotics are all thought to inhibit parasite growth through the inhibition of ‘prokaryote-like’ protein biosynthesis in the apicoplast — an organelle that is unique to apicomplexan parasites such as *Plasmodium* (refs 36,37; and see below).

Mechanisms associated with drug targets

Anti-infective drugs rely for their efficacy and specificity on their ability to interfere with aspects of metabolism that differ significantly from the human host. *Plasmodium* infects host erythrocytes during the phase of their life cycle that gives rise to the symptoms of malaria. Parasite survival in this environment requires several metabolic adaptations and innovations that render it susceptible to chemotherapeutic attack. This is manifest both in the mechanism of action of existing drugs and in the plethora of potential drug targets being identified, which is assisted by data from the malaria genome project (refs 38,39; see also <http://sites.huji.ac.il/malaria/>).

Many drug targets can be related to the functions of distinct organellar structures (ref. 40; and Box 2). Of particular interest are the lysosomal food vacuole (the site of extensive haemoglobin degradation), the apicoplast (a plastid organelle thought to originate from a green algal symbiont) and an acrystate mitochondrion with a limited electron transport system. Other aspects of metabolism are also important and are further highlighted.

Haemoglobin degradation in the food vacuole

During its 48-hour cycle of invasion, growth and release from an infected erythrocyte, the malaria parasite degrades up to 80% of the haemoglobin in the host cell⁴¹. This degradation occurs in a lysosomal food vacuole and involves aspartic proteases (plasmepsins)⁴², the cysteine protease falcipain 2 (ref. 43), and many peptidases including a metallopeptidase⁴⁴. This results in the release of large amounts of Fe(II) haem, which is rapidly oxidized to Fe(III) haematin and sequestered as an inert pigment called haemozoin⁴¹ that comprises a structured lattice of aggregated haem dimers⁴⁵. Both the quinoline and aryl alcohol antimalarials^{46–50} and the artemisinins and other antimalarial peroxides^{17,51} are concentrated in the food vacuole and are thought to exert their activity through interaction with haem.

The quinolines are thought to disrupt or prevent effective formation of haemozoin by binding to haem through π – π stacking of their planar aromatic structures⁵², resulting in haem-mediated toxicity to the parasite. This may occur by inducing lipid peroxidation⁵³, although there are other explanations⁵⁴. The artemisinins undergo oxidoreductive cleavage of their peroxide bond in the food vacuole, most probably through interaction with Fe(II) haem^{17,51}. This generates fatal free-radical-induced damage to the parasite; however, the exact mechanisms by which free radicals are generated and the mechanism of parasite death are still matters of debate⁵⁵.

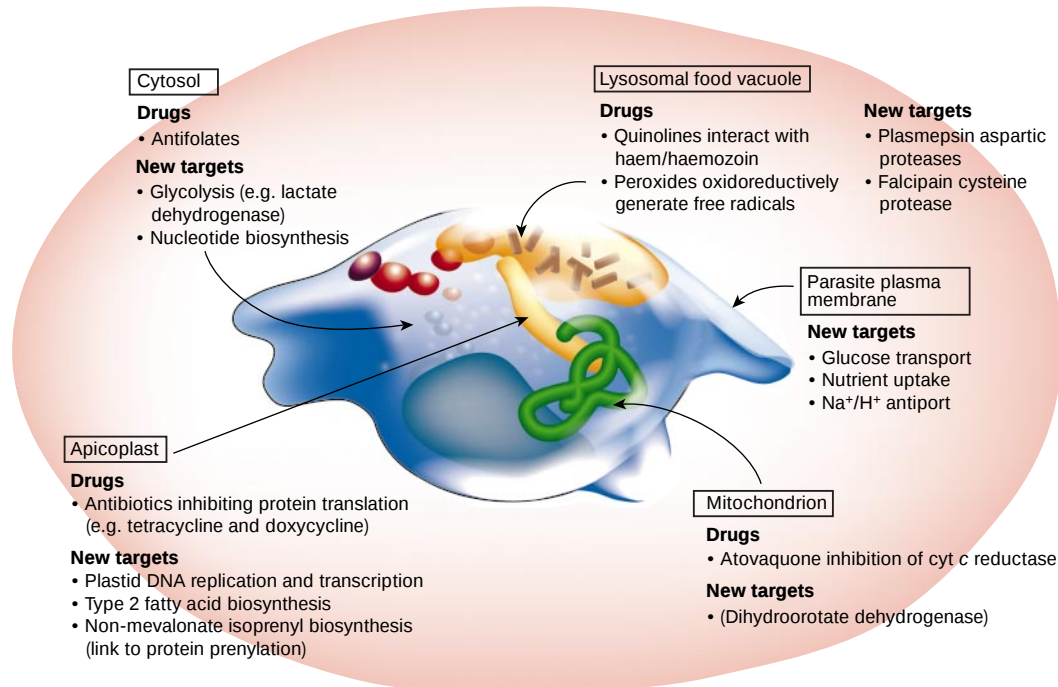
The lack of an enzyme drug target for both quinoline and artemisinin antimalarials is probably a chief reason why resistance development to these drugs is relatively slow^{2,33}. For both classes of drug, resistance may be associated with mutations in transporters that affect drug access to the food vacuole. Data from genetic linkage experiments, transfection experiments and field monitoring of chloroquine-resistant parasites strongly suggest that *P. falciparum* resistance to chloroquine, and potentially to some other 4-aminoquinolines, is linked to the putative chloroquine resistance transporter gene *PfCRT*⁵⁶. But this gene is not associated with chloroquine resistance in *P. vivax*⁵⁷.

A second gene encoding a P-glycoprotein, *mdr1*, has been linked to the altered susceptibility of parasites to mefloquine, halofantrine and the artemisinins in cell-culture studies of a genetic cross⁵⁸ and

Box 2

Sites of drug action and new drug targets

Diagram of *P. falciparum* trophozoite residing in an erythrocyte. The main organelles that are associated with drug targets are highlighted, drawing attention both to sites of current antimalarial drug action and new targets that are under investigation. The concept for this representation is derived from the trophozoite illustrations of Bannister⁴⁰.



through transfection of mutant genes from resistant strains into drug-sensitive parasites⁵⁹. Amplification of this gene may also contribute significantly to resistance to mefloquine and halofantrine⁶⁰. Despite these data, the association between *Pfmdr1* and resistance to mefloquine, halofantrine and quinine remains incomplete, suggesting that other genetic factors, including *PfCRT*, are involved¹³. It should also be noted that although mutations in *Pfmdr1* result in reduced sensitivity of parasites to artemisinin derivatives, no significant clinical drug resistance has been observed against the artemisinins.

Haemoglobin degradation and interaction with haem remain fruitful areas for drug discovery. Relatively small changes in quinoline structure can enhance compound uptake into the food vacuole of chloroquine-resistant parasites^{10,61}. Efforts are currently directed towards short-chain chloroquine analogues, bisquinolines and analogues of amodiaquine that lack the ability to form a toxic metabolite. Significantly, pyronaridine — a 4-aminoquinoline developed originally in China — is undergoing development in combination with artesunate. Molecules that inhibit parasite growth through binding to haem in an analogous way to the binding of quinolines are also being sought through high-throughput screening⁶².

Several molecules have been identified that enhance the activity of chloroquine against chloroquine-resistant strains, most probably by enhancing the uptake of chloroquine into the parasite food vacuole^{13,63,64}. A study in Nigeria in which chloroquine was combined with chlorpheniramine⁶⁵ has suggested that this approach may have clinical application, but further evaluation is needed.

Research chemists are attempting to identify semisynthetic and fully synthetic endoperoxides that will retain the activity of the artemisinin derivatives, but overcome some of their deficiencies, notably their short half-life and the residual fears of potential neurotoxicity^{16,66–69}. Unexploited targets located in the food vacuole include the proteases associated with haemoglobin degradation, namely plasmeypsin aspartic proteases^{41,42} and the cysteine protease falcipain 2 (ref. 43). In both cases, malaria chemotherapy might

benefit from the intense medicinal chemistry efforts under way in the pharmaceutical industry, including the development of inhibitors against HIV aspartic protease and the aspartic protease renin (an enzyme associated with hypertension), and interest in the cysteine proteinase cathepsin K (which is implicated in osteoporosis). The challenge to chemists working in this area is to identify simple, non-peptidic structures with good oral bioavailability and low costs. Evidence that inhibitors of aspartic proteases and cysteine proteases are synergistic in their efficacy against *P. falciparum* in culture provides further opportunity in this area⁴¹.

Apicoplast metabolism and drug targets

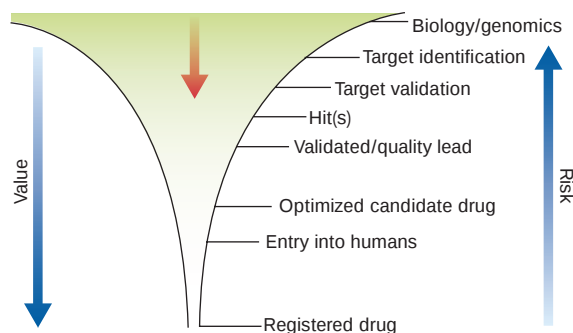
The malaria genome project has made a major contribution to elucidating the metabolism of the apicoplast⁷⁰. This plastid organelle contains a 35-kilobase circular genome that encodes elements of a prokaryotic transcription and translation system⁷¹, suggesting that this organelle is the target of the clinically used antimalarial antibiotics tetracycline, doxycycline and clindamycin, which target prokaryotic translation^{36,37}. Preclinical work has also shown that malaria parasites are susceptible to prokaryotic transcriptional inhibitors such as rifampicin and DNA gyrase inhibitors such as the quinolones^{36,37}.

These antibiotics tend to kill the parasite slowly, but a combination of *Toxoplasma* molecular genetics and *P. falciparum* genomic information has facilitated the discovery of several key metabolic pathways in the apicoplast, the disruption of which results in a more rapid cell death. The first of these is type II fatty acid biosynthesis⁷², a pathway that is often associated with plant eukaryotes and bacteria. Several lead molecules have been identified in this pathway³⁷: thiolactomycin inhibits the condensing enzymes Fab B, Fab F and Fab H in *Escherichia coli* and inhibits *P. falciparum* growth; triclosan is active against parasite growth in cell-culture and rodent models through inhibition of enoyl-acyl-carrier protein reductase⁷³.

Another enzyme pathway identified through genomics and now firmly associated with the apicoplast is the non-mevalonate pathway that leads to synthesis of isopentenyl diphosphate subunits.

Box 3 Drug development funnel

The funnel provides an overview of the process of drug discovery and development. Many possibilities exist in terms of potential targets, but only a limited number of targets are valid and even fewer generate a chemical compound that justifies serious medicinal chemistry. Few compounds achieve drug candidate status and are worthy of testing in humans, and only a few of these are ultimately registered as a drug. As we move down the funnel, the cost of research on an individual project increases because of the more intensive work that is required to determine preclinical and clinical safety and efficacy, and to produce, manufacture and register the drug. Appropriate selection of which projects move forward and which are abandoned is the key to R&D success. As we move down the funnel, the risk of the project failing decreases and its value increases. The cost of an individual project proceeding through this process probably amounts to tens of millions of dollars. Taking into account the number of projects that fail, the cost of a success may be of the order of hundreds of millions of dollars.



Moreover, fosmidomycin, an inhibitor of 1-deoxy-D-xylulose-5-phosphate synthase that was identified in an industrial antibacterial drug discovery programme, inhibits effectively the malarial enzyme and parasite growth⁷⁴, and offers hope either as an antimalarial agent in its own right or as a lead compound. The isopentenyl diphosphate generated by this pathway is also important for protein farnesylation. This is significant because the human protein farnesyltransferases have been investigated by pharmaceutical companies as anticancer drug targets. Recent studies have identified inhibitors of protein farnesyltransferase as potential leads for new antimalarials⁷⁵.

Although the science underlying the enzymes involved in apicoplast metabolism is exciting, it is important to remember that when other enzyme inhibitors, namely pyrimethamine and atovaquone, were used alone in chemotherapy, resistance emerged rapidly. Hopefully, inhibitors of the different pathways of the apicoplast act as synergists in a manner similar to that of the antifolates and atovaquone/proguanil, and lead ultimately to combination treatments of high medical value.

Mitochondrion electron transport

The *Plasmodium* mitochondrion is unusual morphologically⁴⁰. *Plasmodium* parasites reside in an oxygen-deficient environment and rely on glycolysis for ATP production. Thus, the mitochondrion has no oxidative phosphorylation activity and has an incomplete electron transport chain that provides certain redox reactions of metabolic significance. One such reaction is the coupling of cytochrome *c* reductase (cytochrome *b*/cytochrome *c1* complex) to dihydroorotate dehydrogenase, a key enzyme in nucleotide biosynthesis. Atovaquone is known to inhibit electron transport at this point⁷⁶. Similar to experiences with the DHFR inhibitor pyrimethamine, resistance has been found to develop rapidly owing to single-point mutations in cytochrome *b* (ref. 35). This has been

addressed by combining atovaquone with proguanil, which acts in synergy as the prodrug with atovaquone^{34,35}.

Other classes of compound are also known to inhibit cytochrome *c* reductase, notably the β -methoxyacrylates⁷⁷. Compounds from this class are active against atovaquone-resistant strains, but resistance develops very rapidly against them in rodent models of malaria (H. Matile, personal communication).

Folate pathway

The genetic basis of resistance to the antifolate antimalarials is probably the best understood of all the drug-resistance mechanisms described to date^{3,4,78}. High-level pyrimethamine resistance results from the accumulation of mutations in DHFR, principally at codons 108, 59, 51 and 164, where allelic variation gives rise to the mutations Ser108→Asn, Cys59→Arg, Asn51→Ile and Ile164→Leu. Mutations occurring in DHPS at codons 436, 437, 581 and 613 in cultured parasite lines broadly correlate with estimated levels of sulphadoxine resistance. Variability in codon 540 is also observed commonly in field samples. Notably, although uptake of exogenous folate bypasses DHPS, pyrimethamine restores sulphadoxine sensitivity at concentrations much lower than those required to inhibit DHFR, suggesting that pyrimethamine may have additional antiparasitic effects.

Several medicinal chemistry efforts are being directed at identifying improved inhibitors of DHFR that overcome resistance to pyrimethamine. Biguanides related to proguanil, which are metabolized to active triazines, offer the most immediately promising results. Chlorproguanil, which is metabolized to chlorcycloguanil, retains activity against all but the most highly pyrimethamine-resistant strains, and it is anticipated that a fixed-dose combination of this compound with dapsone should achieve registration early in 2002 (ref. 79). Efforts are under way to combine chlorproguanil and dapsone with artesunate to lessen the chances of drug resistance development. Another set of potential biguanide prodrugs also shows promise⁸⁰. Our understanding of how DHFR mutations confer resistance, our ability to model the DHFR active site, and the ease of the DHFR assay suggest that further improved inhibitors are achievable. Many other enzymes in the folate pathway may also be potential drug targets^{38,81}.

Other promising areas for drug discovery

Although it is not possible to discuss all of the many scientific opportunities that are presenting themselves for malaria drug discovery, several other areas deserve to be highlighted. Because malaria parasites are microaerophilic homolactate fermenters and rely on glycolysis for ATP production, lactate dehydrogenase is being explored as a drug target⁸². Because the parasite requires enormous amounts of glucose uptake from the host to support its growth, the identification of a glucose transporter by functional transfection of the gene into *Xenopus* oocytes is potentially significant⁸³.

Other transport studies are further identifying potential drug targets⁸⁴, including an anion-selective channel that is responsible for transporting solutes and nutrients⁸⁵. Proteases involved in erythrocyte invasion also offer potential⁸⁶, as do some highly active compounds that may interfere with phospholipid biosynthesis⁸⁷. Cell signalling pathways, particularly through protein kinases, offer huge potential for drug discovery⁸⁸, and there is a wealth of medicinal chemistry experience available in this area from industry. Until protein kinases can be linked to specific functions, however, it will be difficult to assign a relevance to particular genes for drug discovery.

Power of genomics

The power of genomics and molecular genetics has underpinned many of the biological advances mentioned here, from determining the mechanisms of chloroquine drug resistance to establishing the metabolic pathways associated with the apicoplast. The power that the full malaria genome sequence will bring and the genomic tools at our disposal have been dealt with in depth by others (ref. 89; and see

Hoffman *et al.*, pp. 702–709, in this issue). There is already industrial evidence supporting the power of genomic approaches in microbial drug discovery⁹⁰. Bioinformatics tools can help predict the structure and function of gene products⁹¹. Combined with the power of ‘virtually’ docking compounds into modelled protein structures and of robotics for high-throughput screening, the speed with which many compounds can be tested against a protein target is ceasing to be a limiting factor.

For malaria, the main bottleneck in the functional genomics of *P. falciparum* is the limitation of our transfection systems, although many examples of successful transfection experiments now exist^{89,92} and improvements are being made continuously. Currently, it takes about 3–4 weeks to obtain a single transfectant owing to low transfection rates and the (relatively) slow growth of the parasite. Only single crossover insertions have been observed so far in *P. falciparum*⁹². Transfection capabilities are also available for some model animal systems, including *Plasmodium berghei*^{89,93}. In many cases, the enhanced efficiencies of these processes allow double crossovers and gene replacement to occur.

We also gain insight into *P. falciparum* biology by using molecular genetics tools developed for *Toxoplasma gondii*, a related apicomplexan parasite⁷² that has the additional advantage of allowing non-homologous integration. The ability to perform knockout experiments more routinely and gene-replacement experiments on a larger scale in *P. falciparum* would advance the field greatly. Of particular importance for drug discovery is the development of inducible knockout techniques, such as tetracycline-inducible systems and RNA inhibition, that can transiently and selectively knockout gene function to allow observation of the resultant phenotype.

Transfection technology will produce numerous tools to assist drug discovery and development. The replacement of a gene encoding a drug target in *P. berghei* with its *P. falciparum* homologue is currently being developed (D. Fidock and W. Jacobs, personal

communication) and deserves comment. A chief difficulty in drug discovery is moving from molecular assays, to cellular assays, to appropriate animal models. Unfortunately for malaria drug discovery, the human malaria parasites do not infect rodents, and specific rodent malaria parasites such as *P. berghei* have to be used. For targets associated with haem, such as the quinolines and the peroxides, the rodent malaria parasite *P. berghei* is an excellent model. But for enzyme targets in which there may be differences in protein sequence, structure and inhibitor specificity, this ceases to be the case. In the absence of sufficient similarity between the enzymes of *P. berghei* and *P. falciparum*, one has either to rely on cell-culture data combined with rapid pharmacokinetic evaluation or to move to much more expensive (and low-throughput) monkey models that can sustain infection by *P. falciparum*. The ability to test inhibitors against a *P. falciparum* (or *P. vivax*) enzyme in a rodent system would greatly facilitate rapid compound evaluation and feedback to medicinal chemists on aspects such as oral bioavailability and efficacy.

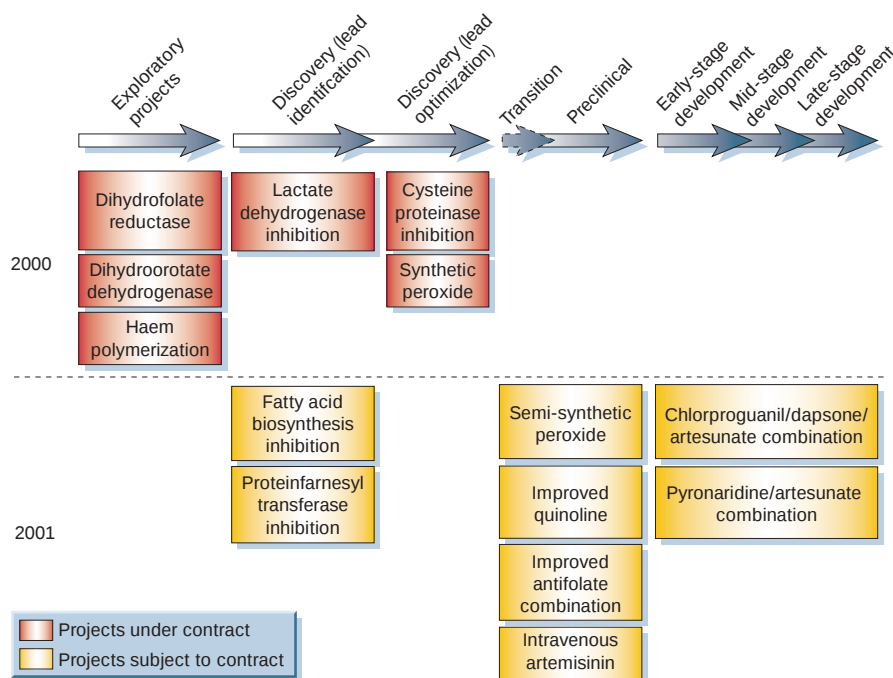
Focused drug discovery and development

When discussing issues relating to drug discovery and development, it is necessary to keep in mind the many disciplines and resources that have to come together to deliver success⁹⁴. This should be coupled with the understanding that success is by no means guaranteed. Box 3 shows a ‘funnel’ model of the discovery and development process. Genomics can provide many potential molecular targets, but only those targets that can be readily manipulated and tested are likely to provide an opportunity to discover a lead molecule, for example by high-throughput screening, that is worth taking on to a full medicinal chemistry project. Thus, many potential targets will never be advanced because no chemistry lead is identified. Once a chemistry effort is started, many factors need to be considered over and above improving enzyme inhibition and efficacy against the parasite in culture. The molecules finally selected for development must also be

Box 4

Significance of developing a portfolio of projects

An example of a drug research and development portfolio taken from the 2001 Medicines for Malaria Venture annual report (<http://www.mmv.org>). The arrows refer to the discovery and development process that takes about 10–15 years. It comprises an exploratory phase in which a project attempts to develop the tools and lead molecules necessary for a concerted discovery effort; a discovery phase in which lead compounds are better characterized, developed and optimized using medicinal chemistry; a transition and preclinical development phase in which candidate compounds undergo extensive analysis with respect to toxicity, adsorption, pharmacokinetics and formulation; and finally the various clinical phases during which continued manufacturing, production and toxicology studies are undertaken. Some of the projects for 2001 are still subject to contract. The Medicines for Malaria Venture has calculated that it needs to double the current portfolio to ensure the registration of one new drug every 5 years.



easy to manufacture (low cost is crucial for antimalarials), stable, readily formulated, bioavailable (that is, extensively adsorbed from the gut and avoiding first pass metabolism in the liver to achieve effective concentrations in the systemic circulation), have an appropriate half-life, and not show any overt toxicity.

It is essential that academic scientists better understand these issues as they are increasingly having to take the lead in drug discovery efforts against malaria and other neglected diseases. An analysis of the reasons why candidate drugs fail to achieve registration and reach the market has been undertaken recently⁹⁵. In 39% of cases, drugs failed after entering development because of 'biopharmaceutical' issues such as oral bioavailability and formulation, and in 21% of cases because of toxicity. These issues are equally as important as drug efficacy, for which 29% of cases failed.

Over the past 5 years, progress has been made in defining rules to assist chemists in the synthesis of compounds that will possess adequate properties for formulation and bioavailability. This is perhaps best embodied in variations of Lipinski's rules⁹⁶, which define appropriate parameters for optimizing solubility and membrane permeability such as log *P* (where *P* is the partition coefficient between octanol and water), molecular weight, and the number of hydrogen bond donor and/or acceptor groups. Appropriate modelling and analysis of compounds for these parameters is ultimately as important to obtaining a drug as the molecular modelling of compounds into the active sites of enzymes. Information technology and sophisticated databases also enable chemists to identify molecular groups that are likely to be metabolically labile or that have associated toxicity.

Successful drug discovery and development requires manpower, financial resources and management. The skills required and the experience of bringing these areas of expertise together in a team exist primarily in the pharmaceutical industry. In recognition of this key industrial role, the Medicines for Malaria Venture was established (<http://www.mmv.org>). This organization builds on the experiences of the Special Program for Research and Training in Tropical Diseases (<http://www.who.int/tdr>) and operates in concert with the Roll Back Malaria campaign of the WHO (<http://www.rbm.who.int>). It seeks through appropriate funding to bring together teams of scientists from both academic and industrial backgrounds to form powerful drug discovery and development project teams. The projects that are funded combine to generate a portfolio (Box 4) that over time is expected to deliver new antimalarials at regular and optimum intervals.

Current drug discovery and development efforts centre on identifying molecules that will be active in treating uncomplicated malaria. The ideal product profile comprises orally active compounds that can cure the disease with a 3-day regimen using once-a-day dosing. As a strong portfolio develops, however, it is anticipated that other, more specific malarial indications may be targeted, such as adjunct treatments to improve the outcome of severe malaria cases; intermittent treatment of malaria during pregnancy to protect both the mother from the disease and the unborn child from a higher risk of being born underweight³¹; and long half-life drugs to treat malaria with a single dose in complex emergency situations. There is also the prospect of combining compounds from several projects into combination products. This could both enhance efficacy and reduce the likelihood of drug resistance development.

Drug access

Good drugs are only useful if they are made available to the people who need them and are used properly. Thus, the availability of inexpensive compounds, sound policies on their use, strong healthcare infrastructure, and sufficient funds to purchase the drugs that are needed are all vital^{5,79}. The political will and momentum that are being generated in the countries worst affected by the disease through the WHO-led global campaign to Roll Back Malaria⁹⁷, combined with renewed efforts to obtain both sufficient global support and

funds to purchase drugs⁹⁸ and the increased science funding going into malaria drug discovery and development, should result in a significant reduction in the malaria disease burden over the coming years and decades.

Progress in drug discovery and development should deliver a larger choice of drugs to treat malaria. This will reduce a current dilemma that faces many malaria control agencies. Because of the lack of alternatives they find it difficult to balance the need to distribute the most effective drugs widely against the need to minimize resistance development through controlled use. The world needs to reach a situation in which many effective, appropriate and affordable antimalarial drugs are in reserve to meet the challenge of resistance as and when it arises. □

- Calabrese, L. & Fleischer, A. Thalidomide: current and potential clinical applications. *Am. J. Med.* **108**, 487–495 (2000).
- Wellems, T. & Plowe, C. Chloroquine-resistant malaria. *J. Infect. Dis.* **184**, 770–776 (2001).
- Sirawaraporn, W. Dihydrofolate reductase and antifolate resistance in malaria. *Drug Resist. Update* **1**, 397–406 (1998).
- Cowman, A. in *Malaria Parasite Biology, Pathogenesis and Protection* (ed. Sherman, I.) 317–330 (ASM, Washington, 1999).
- Report No. WHO/CDS/RBM/2001.33 (World Health Organization, Geneva, 2001).
- Report No. WHO/CDS/RBM/2001.35 (World Health Organization, Geneva, 2001).
- Report No. WHO/MAL/98.1086 (World Health Organization, Geneva, 1998).
- Rosenthal, P. (ed.) *Antimalarial Chemotherapy. Mechanisms of Action, Resistance, and New Directions in Drug Discovery* (Humana, Totowa, New Jersey, 2001).
- Meshnick, S. & Dobson, M. in *Antimalarial Chemotherapy. Mechanisms of Action, Resistance, and New Directions in Drug Discovery* (ed. Rosenthal, P.) 15–26 (Humana, Totowa, New Jersey, 2001).
- Ridley, R. G. & Hudson, A. T. Quinoline anti-malarials. *Exp. Opin. Ther. Patents* **8**, 121–136 (1998).
- O'Neill, P., Bray, P., Hawley, S., Ward, S. & Park, B. 4-Aminoquinolines—past, present, and future: a chemical perspective. *Pharmacol. Ther.* **77**, 29–58 (1998).
- Tilley, L., Loria, P. & Foley, M. in *Antimalarial Chemotherapy. Mechanisms of Action, Resistance, and New Directions in Drug Discovery* (ed. Rosenthal, P.) 87–122 (Humana, Totowa, New Jersey, 2001).
- Dorsey, G., Fidock, D., Wellems, T. & Rosenthal, P. in *Antimalarial Chemotherapy. Mechanisms of Action, Resistance and New Directions in Drug Discovery* (ed. Rosenthal, P.) 153–172 (Humana, Totowa, New Jersey, 2001).
- Li, Y. & Wu, Y.-L. How chinese scientists discovered qinghaosu (artemisinin) and developed its derivatives. What are the future perspectives? *Med. Trop.* **58**, 9–12 (1998).
- Price, R. Artemisinin drugs: novel antimalarial agents. *Exp. Opin. Invest. Drugs* **9**, 1815–1827 (2000).
- Haynes, R. Artemisinin and derivatives: the future for malaria treatment? *Curr. Opin. Infect. Dis.* **14**, 719–726 (2001).
- Meshnick, S. in *Antimalarial Chemotherapy. Mechanisms of Action, Resistance, and New Directions in Drug Discovery* (ed. Rosenthal, P.) 191–202 (Humana, Totowa, New Jersey, 2001).
- White, N. Assessment of the pharmacodynamic properties of antimalarial drugs in vivo. *Antimicrob. Agents Chemother.* **41**, 1413–1422 (1997).
- White, N. Delaying antimalarial drug resistance with combination chemotherapy. *Parasitologia* **41**, 301–308 (1999).
- Brockman, A. et al. *Plasmodium falciparum* antimalarial drug susceptibility on the north-western border of Thailand during five years of extensive use of artesunate-mefloquine. *Trans. R. Soc. Trop. Med. Hyg.* **94**, 537–544 (2000).
- Bloland, P. B., Ertling, M. & Meek, S. Combination therapy for malaria in Africa: hype or hope? *Bull. World Health Organ.* **78**, 1378–88 (2000).
- Petrus, J. et al. Arteether-induced brain injury in *Macaca mulatta*. I. The precerebellar nuclei: the lateral reticular nuclei, paramedian reticular nuclei, and perihypoglossal nuclei. *Anat. Embryol. (Berl.)* **201**, 383–397 (2000).
- Dayan, A. Neurotoxicity and artemisinin compounds. Do the observations in animals justify limitation of clinical use? *Med. Trop.* **58**, 32–37 (1998).
- Kissinger, E. et al. Clinical and neurophysiological study of the effects of multiple doses of artemisinin on brain-stem function in Vietnamese patients. *Am. J. Trop. Med. Hyg.* **63**, 2000 (2000).
- Leonardi, E., Gilvary, G., White, N. & Nosten, F. Severe allergic reactions to oral artesunate: a report of two cases. *Trans. R. Soc. Trop. Med. Hyg.* **95**, 182–183 (2001).
- Skelton-Stroud, P. & Mull, R. Positioning, labelling, and medical information control of co-artemether tablets (CPG 56697): a fixed novel combination of artemether and benflumetol. *Med. Trop.* **58**, 77–81 (1998).
- Noeld, H., Allmendinger, T., Prajakwong, S., Wernsdorfer, G. & Wernsdorfer, W. Desbutyl-benflumetol, a novel antimalarial compound: in vitro activity in fresh isolates of *Plasmodium falciparum* from Thailand. *Antimicrob. Agents Chemother.* **45**, 2106–2109 (2001).
- Basco, L., Bickel, J. & Ringwald, P. In vitro activity of lumefantrine (benflumetol) against clinical isolates of *Plasmodium falciparum* in Yaounde, Cameroon. *Antimicrob. Agents Chemother.* **42**, 2347–2351 (1998).
- Ezzet, F., Vugt, M. v., Nosten, F., Looareesuwan, S. & White, N. Pharmacokinetics and pharmacodynamics of lumefantrine (benflumetol) in acute falciparum malaria. *Antimicrob. Agents Chemother.* **44**, 697–704 (2000).
- Sherman, I. in *Malaria—Parasite Biology, Pathogenesis and Protection* (ed. Sherman, I.) 177–184 (ASM, Washington DC, 1998).
- Wolfe, E. et al. Cost-effectiveness of sulfadoxine-pyrimethamine for the prevention of malaria-associated low birth weight. *Am. J. Trop. Med. Hyg.* **64**, 178–186 (2001).
- Takechi, M. et al. Therapeutic efficacy of sulphadoxine-pyrimethamine and susceptibility in vitro of *P. falciparum* isolates to sulphadoxine-pyrimethamine and other antimalarial drugs in Malawian children. *Trop. Med. Int. Health* **6**, 429–434 (2001).
- White, N. Drug resistance in malaria. *Br. Med. Bull.* **54**, 703–715 (1998).

34. Looareesuwan, S., Chulay, J., Canfield, C. & Hutchinson, D. Malarone (atovaquone and proguanil hydrochloride): a review of its clinical development for treatment of malaria. *Malarone Clinical Trials Study Group. Am. J. Trop. Med. Hyg.* **60**, 533–541 (1999).
35. Vaidya, A. in *Antimalarial Chemotherapy. Mechanisms of Action, Resistance, and New Directions in Drug Discovery* (ed. Rosenthal, P.) 203–218 (Humana, Totowa, New Jersey, 2001).
36. Fichera, M. & Roos, D. A plastid organelle as a drug target in apicomplexan parasites. *Nature* **390**, 407–409 (1997).
37. Clough, B. & Wilson, R. in *Antimalarial Chemotherapy. Mechanisms of Action, Resistance, and New Directions in Drug Discovery* (ed. Rosenthal, P.) 265–286 (Humana, Totowa, New Jersey, 2001).
38. Macreadie, I., Ginsburg, H., Sirawaraporn, W. & Tilley, L. Antimalarial drug development and new targets. *Parasitol. Today* **16**, 438–444 (2000).
39. Padmanaban, G. & Rangarajan, P. Emerging targets for antimalarial drugs. *Expert Opin. Ther. Targets* **5**, 423–441 (2001).
40. Bannister, L. A brief illustrated guide to the ultrastructure of *Plasmodium falciparum* asexual blood stages. *Parasitol. Today* **16**, 427–433 (2000).
41. Francis, S., Sullivan, D. J. & Goldberg, D. Hemoglobin metabolism in the malaria parasite *Plasmodium falciparum*. *Annu. Rev. Microbiol.* **51**, 97–123 (1997).
42. Coombs, G. et al. Aspartic proteases of *Plasmodium falciparum* and other parasitic protozoa as drug targets. *Trends Parasitol.* **17**, 532–537 (2001).
43. Shenai, B., Sijwali, P., Singh, A. & Rosenthal, P. Characterization of native and recombinant falcipain-2, a principal trophozoite cysteine protease and essential hemoglobinase of *Plasmodium falciparum*. *J. Biol. Chem.* **275**, 29000–29010 (2000).
44. Eggleston, K., Duttin, K. & Goldberg, D. Identification and characterisation of Falcilysin, a metalloprotease involved in hemoglobin catabolism within the malarial parasite *Plasmodium falciparum*. *J. Biol. Chem.* **274**, 32411–32417 (1999).
45. Pagola, S., Stephens, P., Böhle, D., Kosar, A. & Madsen, S. The structure of malaria pigment β -haematin. *Nature* **16**, 307–310 (2000).
46. Chou, A., Cherli, R. & Fitch, C. Ferriprotoporphyrin IX fulfills the criteria for identification as the chloroquine receptor of malaria parasites. *Biochemistry* **19**, 1543–1549 (1980).
47. Sullivan, D. J., Gluzman, I., Russell, D. & Goldberg, D. On the molecular mechanism of chloroquine's antimalarial action. *Proc. Natl Acad. Sci. USA* **93**, 11865–11870 (1996).
48. Bray, P., Munghin, M., Ridley, R. & Ward, S. Access to hemozoin: the basis of chloroquine resistance. *Mol. Pharmacol.* **54**, 170–179 (1998).
49. Munghin, M., Bray, P., Ridley, R. & Ward, S. Central role of hemoglobin degradation in mechanisms of action of 4-aminoquinolines, quinoline methanols, and phenanthrene methanols. *Antimicrob. Agents Chemother.* **42**, 2973–2977 (1998).
50. Sullivan, D. J., Matile, H., Ridley, R. & Goldberg, D. A common mechanism for blockade of heme polymerization by antimalarial quinolines. *J. Biol. Chem.* **273**, 31103–31107 (1998).
51. Meshnick, S., Taylor, T. & Kamchonwongpaisan, S. Artemisinin and the antimalarial endoperoxides: From herbal remedy to targeted chemotherapy. *Microbiol. Rev.* **60**, 301–315 (1996).
52. Ridley, R., Dorn, A., Vippagunta, S. & Vennerstrom, J. Haematin (haem) polymerisation and its inhibition by quinoline antimalarials. *Ann. Trop. Med. Parasitol.* **91**, 559–566 (1997).
53. Sugioaka, Y. & Suzuki, M. The chemical basis for the ferriprotoporphyrin IX–chloroquine complex induced lipid peroxidation. *Biochim. Biophys. Acta* **1074**, 19–24 (1991).
54. Ginsburg, H. & Kruglik, M. Chloroquine—some open questions on its antimalarial mode of action and resistance. *Drug Resist. Update* **2**, 63–70 (1999).
55. Olliaro, P., Haynes, R., Meunier, B. & Yuthavong, Y. Possible modes of action of the artemisinin-type compounds. *Trends Parasitol.* **17**, 122–126 (2001).
56. Fidock, D. A. et al. Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol. Cell* **6**, 861–871 (2000).
57. Nomura, T. et al. Evidence for different mechanisms of chloroquine resistance in 2 *Plasmodium* species that cause human malaria. *J. Infect. Dis.* **183**, 1653–1661 (2001).
58. Duraisingh, M., Roper, C., Walliker, D. & Warhurst, D. Increased sensitivity to the antimalarials mefloquine and artemisinin is conferred by mutations in the *pfmdr1* gene of *Plasmodium falciparum*. *Mol. Microbiol.* **36**, 955–961 (2000).
59. Reed, M. B., Saliba, K. J., Caruana, S. R., Kirk, K. & Cowman, A. F. Pgh1 modulates sensitivity and resistance to multiple antimalarials in *Plasmodium falciparum*. *Nature* **403**, 906–909 (2000).
60. Cowman, A., Galatis, D. & Thompson, J. Selection for mefloquine resistance in *Plasmodium falciparum* is linked to amplification of the *pfmdr1* gene and cross-resistance to halofantrine and quinine. *Proc. Natl Acad. Sci. USA* **91**, 1143–1147 (1994).
61. Stocks, P. A., Raynes, K. & Ward, S. in *Antimalarial Chemotherapy. Mechanisms of Action, Resistance, and New Directions in Drug Discovery* (ed. Rosenthal, P.) 235–253 (Humana, Totowa, New Jersey, 2001).
62. Kurosawa, Y. et al. Hematin polymerization assay as a high-throughput screen for identification of new antimalarial pharmacophores. *Antimicrob. Agents Chemother.* **44**, 2638–2644 (2000).
63. Martin, S., Oduola, A. & Milhous, W. Reversal of chloroquine resistance in *Plasmodium falciparum* by verapamil. *Science* **235**, 899–901 (1987).
64. Bray, P. & Ward, S. A comparison of the phenomenology and genetics of multidrug resistance in cancer cells and quinoline resistance in *Plasmodium falciparum*. *Pharmacol. Ther.* **77**, 1–28 (1998).
65. Sowunmi, A. et al. Enhanced efficacy of chloroquine-chlorpheniramine combination in acute uncomplicated falciparum malaria in children. *Trans. R. Soc. Trop. Med. Hyg.* **91**, 63–67 (1997).
66. Posner, G. Antimalarial peroxides in the qinghaosu (artemisinin) and yinghaosu families. *Expert Opin. Ther. Patents* **8**, 1487–1493 (1998).
67. Expert Opinion on: Water soluble trioxanes as potent and safe antimalarial agents. *Exp. Opin. Ther. Patents* **11**, 1351–1354 (2001).
68. Vroman, J., Alvim-Gaston, M. & Avery, M. Current progress in the chemistry, medicinal chemistry and drug design of artemisinin-based antimalarials. *Curr. Pharmaceut. Des.* **5**, 101–108 (1999).
69. Dong, Y. & Vennerstrom, J. Peroxidic antimalarials. *Exp. Opin. Ther. Patents* **11**, 1753–1760 (2001).
70. Kohler, S. et al. A plastid of probable green algal origin in Apicomplexan parasites. *Science* **275**, 1485–1489 (1997).
71. Wilson, R. et al. Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. *J. Mol. Biol.* **261**, 155–172 (1996).
72. Waller, R. et al. Nuclear encoded proteins target the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **95**, 12352–12357 (1998).
73. Surolia, N. & Surolia, A. Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of *Plasmodium falciparum*. *Nature Med.* **7**, 167–173 (2001).
74. Jomaa, H. et al. Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science* **285**, 1573–1576 (1999).
75. Ohkanda, J. et al. Peptidomimetic inhibitors of protein farnesyltransferases show antimalarial activity. *Bioorg. Med. Chem. Lett.* **11**, 761–764 (2001).
76. Fry, M. & Beesley, J. Mitochondria of mammalian *Plasmodium* spp. *Parasitol. Today* **102**, 17–26 (1991).
77. Alzeer, J. et al. Phenyl beta-methoxyacrylates: a new antimalarial pharmacophore. *J. Med. Chem.* **43**, 560–568 (2000).
78. Hyde, J. Mechanisms of resistance of *Plasmodium falciparum* to antimalarial drugs. *Microb. Infect.* (in the press).
79. Winstanley, P. Chemotherapy for falciparum malaria: the armoury, the problems and the prospects. *Parasitol. Today* **16**, 146–153 (2000).
80. Kinyanjui, S., Mberu, E., Winstanley, P., Jacobus, D. & Watkins, W. The antimalarial triazine WR99210 and the prodrug PS-15: folate reversal of in vitro activity against *Plasmodium falciparum* and a non-antifolate mode of action of the prodrug. *Am. J. Trop. Med. Hyg.* **60**, 943–947 (1999).
81. Olliaro, P. & Yuthavong, Y. An overview of chemotherapeutic targets for antimalarial drug discovery. *Pharmacol. Ther.* **81**, 91–110 (1999).
82. Dunn, C. et al. The structure of lactate dehydrogenase from *Plasmodium falciparum* reveals a new target for anti-malarial design. *Nature Struct. Biol.* **3**, 912–915 (1996).
83. Woodrow, C., Penny, J. & Krishna, S. Intraerythrocytic *Plasmodium falciparum* expresses a high affinity facilitative hexose transporter. *J. Biol. Chem.* **274**, 7272–7277 (1999).
84. Kirk, K. Membrane transport in the malaria-infected erythrocyte. *Physiol. Rev.* **81**, 495–537 (2001).
85. Desai, S., Bezrukov, S. & Zimmerberg, J. A voltage-dependent channel involved in nutrient uptake by red blood cells infected with the malaria parasite. *Nature* **406**, 1001–1005 (2000).
86. Blackman, M. Proteases involved in erythrocyte invasion by the malaria parasite: function and potential as drug targets. *Curr. Drug Targets* **1**, 59–83 (2000).
87. Vial, H. et al. Transport of phospholipid synthesis precursors and lipid trafficking into malaria-infected erythrocytes. *Novartis Found. Symp.* **226**, 82–88 (1999).
88. Kappes, B., Doerrig, C. & Graeser, R. An overview of *Plasmodium* protein kinases. *Parasitol. Today* **15**, 449–454 (1999).
89. Horrocks, P., Bowman, S., Kyes, S., Waters, A. & Craig, A. Entering the post-genomic era of malaria research. *Bull. World Health Organ.* **78**, 1424–1437 (2000).
90. Buysse, J. The role of genomics in antibacterial target discovery. *Curr. Med. Chem.* **8**, 1763–1776 (2001).
91. Teichman, S., Chothia, C. & Gerstein, M. Advances in structural genomics. *Curr. Opin. Struct. Biol.* **9**, 390–399 (1999).
92. Wu, Y., Kirkman, L. & Welles, T. Transformation of *Plasmodium falciparum* malaria parasites by homologous integration of plasmids that confer resistance to pyrimethamine. *Proc. Natl Acad. Sci. USA* **93**, 1130–1134 (1996).
93. Waters, A. et al. Transfection of malaria parasites. *Methods* **13**, 134–147 (1997).
94. Ridley, R. Antimalarial drug discovery and development — an industrial perspective. *Exp. Parasitol.* **87**, 293–304 (1997).
95. Venkatesh, S. & Lipper, R. Role of the development scientist in compound lead selection and optimization. *J. Pharm. Sci.* **89**, 145–154 (2000).
96. Lipinski, C., Lombardo, F., Dominy, B. & Feeney, P. Experimental and computational approaches to estimate solubility in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25 (1997).
97. Alnwick, D. Roll back malaria—what are the prospects? *Bull. World Health Organ.* **78**, 1377 (2000).
98. Sachs, J. A new global commitment to disease control in Africa. *Nature Med.* **7**, 521–523 (2001).

Acknowledgements

I thank W. Jacobs and D. Fidock for allowing me to describe their concept for transfecting genes from human *Plasmodium* species into rodent *Plasmodium* species; T. Sukwa for discussions; and D. Fidock for reading the manuscript.

Progress and challenges for malaria vaccines

Thomas L. Richie* & Allan Saul†

*Malaria Program, Naval Medical Research Center, 503 Robert Grant Avenue, Silver Spring, Maryland 20910-7500, USA
(e-mail: richiet@nmrc.navy.mil)

†Malaria Vaccine Development Unit, NIAID, National Institutes of Health, Twinbrook I, 5640 Fisher Lane, Rockville, Maryland 20853, USA
(e-mail: asaul@niaid.nih.gov)

Malaria causes much physical and economic hardship in tropical regions, particularly in communities where medical care is rudimentary. Should a vaccine be developed, it is the residents of these areas that stand to benefit the most. But the vaccine, which has been promised to be 'just round the corner' for many years, remains elusive. It is important to ask why this is so, when effective vaccines exist for many other infectious diseases. What are the reasons for the slow rate of progress, and what has been learned from the first clinical trials of candidate malaria vaccines? What are the remaining challenges, and what strategies can be pursued to address them?

Four species of malaria infect humans, which raises an initial question of how many vaccines are needed. Studies in the 1950–60s of sequential heterologous infections in malaria therapy patients, such as *Plasmodium vivax* after *Plasmodium falciparum*¹, and cross-species challenge experiments in the 1970s using the irradiated sporozoite vaccine² showed that protection against several species will be difficult to achieve with a single vaccine. Accordingly, species-specific vaccines for *P. falciparum* and *P. vivax* — the two parasites that contribute most heavily to the malaria burden — are being developed, keeping open the possibility of combining the antigens into a single formulation sometime in the future. Once successful vaccines have been developed against *P. falciparum* and *P. vivax*, it should be relatively straightforward to create similar vaccines for *Plasmodium malariae* and *Plasmodium ovale*.

The *P. falciparum* parasite deserves particular attention because of the variety and severity of disease syndromes that it causes (see review in this issue by Miller *et al.*, pages 673–679). Several risk groups are found among those living in endemic areas who are subject to repeated *P. falciparum* infections: infants and young children suffer particularly from life-threatening anaemia, older children from an induced coma³, and primigravida women from severe disease related to placental sequestration⁴. For each of these groups, an anti-morbidity vaccine, possibly tailored to the underlying pathophysiology, would be of great benefit. Malaria-naïve travellers, either crossing international borders or travelling from malaria-free to malaria-endemic areas in their own countries, constitute another risk group, and are susceptible to severe disease after acquiring their first infection; for this group, it is important to prevent malaria infection altogether.

Although practical considerations of both development and production costs favour a single vaccine for *P. falciparum*, the different risk groups and vaccine requirements have generated at least three approaches for this species alone (Box 1): an anti-infection vaccine aimed at protecting malaria-naïve travellers or residents of low endemic areas from becoming infected; an anti-disease/anti-mortality

vaccine aimed at children, pregnant women and migrants living in endemic areas; and an anti-mosquito-stage vaccine aimed at preventing the transmission of malaria from one person to another⁵.

In this review, we discuss the challenges faced in developing these vaccines, why we believe that success is likely, the strategies being pursued and the progress to date, including some encouraging results from clinical trials. We outline the tasks that remain and how they are being addressed by a worldwide effort.

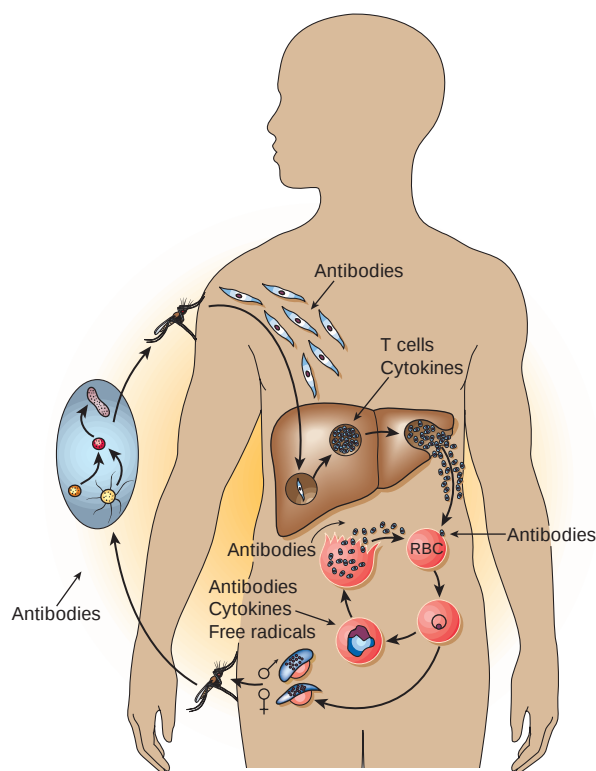
Attacking a chronic infection

To appreciate the constraints of curtailing malaria infection with a vaccine, consider that the parasite that can establish a chronic infection in an immunocompetent host, successfully evading all branches of the immune system. This contrasts with most acute infectious diseases, for which it is relatively straightforward to reproduce the sterile immunity that follows natural infection. There are several immune responses that restrict parasite growth, but the parasite persists. Indeed, the parasite benefits from immune responses if they lead to chronic infection and thereby enhance transmission to the mosquito. As yet, none of the immune responses identified in humans robustly predicts protection from infection or from disease, nor do we know for sure which branches of the immune system will succeed ultimately in eradicating the parasite. Thus, the plan to harness immune responses to destroy the parasite presents a conundrum and requires clever approaches that short-circuit the mechanisms of immune evasion.

Complicating the picture, the different stages of the parasite express different antigens, and a vaccine effective in killing liver-stage parasites may not inhibit the growth of blood-stage parasites. Many parasite proteins exhibit polymorphism, which potentially limits the effectiveness of any vaccine not incorporating distinct variants of antigen. For example, a single parasite clone contains roughly 50 different copies of the gene for the variable surface antigen PfEMP1 (*P. falciparum* erythrocyte membrane protein 1). During a chronic infection, each successive wave of parasitaemia expresses a new variant surface antigen, thus

Box 1

Parasite stages under attack



Current approaches to malaria vaccine development can be classified according to the parasite stages that are targeted:

1. Vaccines directed against sporozoites and/or liver stages (collectively termed pre-erythrocytic stages) are designed to prevent blood-stage infection and thereby avoid all manifestations of disease (anti-infection vaccines).
2. Vaccines directed against asexual blood stages are designed to reduce clinical severity (anti-morbidity/mortality vaccines).
3. Vaccines directed against mosquito stages are designed to halt development in the mosquito (transmission-blocking vaccines). Protective mechanisms of immunity are shown for each stage.

In reality, the effects that can be anticipated for each type of vaccine overlap broadly: a pre-erythrocytic-stage vaccine, even if not 100% effective, could reduce transmission and morbidity — the latter is predicted by the reductions in morbidity and mortality associated with the use of insecticide-impregnated bednets; a highly effective blood-stage vaccine could eliminate blood stages as soon as they emerge from the liver, thereby curtailing both infection and transmission; and a transmission-blocking vaccine could reduce population-wide malaria infection rates and malaria-associated morbidity.

allowing parasite multiplication despite the presence of antibodies directed against the preceding parasitic wave⁶. The sheer number of malaria proteins — estimated to be at least 5,000–6,000 — presents a perplexing choice for the designer of a subunit vaccine (see review in this issue by Hoffman *et al.*, pages 702–709). On the other side of the parasite–host equation, humans exhibit striking heterogeneity in their immune response, depending on their type of human leukocyte antigen (HLA) and other factors (Fig. 1). There is also a bewildering array of other genetic traits, such as haemoglobin type and red cell polymorphisms, that control susceptibility to and have been

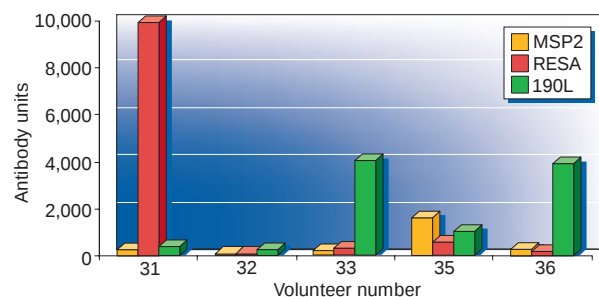


Figure 1 Diversity of immune responses in humans to malaria antigens. Data from five consecutive volunteers (volunteer numbers 31–36; 34 was a placebo) receiving a mixture of 50 µg of each of three malaria antigens, the 190L fragment of MSP1, MSP2 and RESA, in Montanide ISA720, followed by a boost of 20 µg of each antigen³⁰. A similar diversity of responses was seen in volunteers who received the three antigens separately. Tasks for vaccine developers include increasing the immunogenicity of malaria antigens through new formulations and approaches, and designing vaccines that induce protective immune responses in everyone, despite diverse genetic backgrounds.

selected by malaria⁷. Given these complexities, is it possible to make an effective vaccine?

Why a vaccine is feasible

Several studies have shown that malaria vaccines are feasible. First, immunization with irradiated sporozoites protects or partially protects rodents⁸, monkeys⁹ and humans^{2,10,11} from being infected by sporozoites. In humans, immunization through the bites of more than 1,000 irradiated, infected mosquitoes, in batches of a few hundred bites over several months, confers sterilizing protection against sporozoite challenge in more than 90% of vaccine recipients. Although the immune mechanisms underlying this protection are not known, evidence indicates the importance of T-cell responses directed against parasite proteins expressed on the surface of infected hepatocytes¹².

So far we have not identified a single dominant immune response in irradiated sporozoite-immunized hosts¹³. This suggests, but does not establish, that protection may be mediated by the summation of many modest immune responses against the large variety of antigens presented by this attenuated-organism vaccine. Although irradiated sporozoites delivered by mosquito bite are unlikely to be a practical vaccine, their effectiveness indicates that a multivalent pre-erythrocytic (sporozoite and liver)-stage vaccine might induce sterile immunity. This type of anti-infection vaccine would prevent all manifestations of disease and would be especially suited for travellers or others exposed to relatively low intensity transmission, but may also have an important role in limiting malaria-related morbidity and mortality in those living in endemic areas.

Second, people infected repeatedly by malaria develop ‘naturally acquired immunity’ (NAI), which protects against clinical disease¹⁴. But if their parasites are eliminated through radical drug cure, these individuals can become re-infected, which indicates that NAI does not include absolute anti-infection immunity¹⁵. If we could reproduce NAI with a vaccine (that is, use a vaccine to hasten the transition from malaria-naïve to clinically immune), we could shelter recipients from malarial disease. Antibody responses to blood-stage parasites constitute part of NAI, because passive transfer of purified immunoglobulins from immune individuals protects children^{16–18}. Thus, another avenue for designing a vaccine is to generate antibody responses to blood-stage parasite antigens. This type of vaccine, which would suppress blood-stage parasites and prevent disease, would be especially suited to those exposed to intense

transmission and could be used either alone or as a complement to a pre-erythrocytic-stage vaccine. The efficacy of such a vaccine for particular target groups might be improved by tailoring its composition to the underlying pathophysiology, such as the inclusion of components that block placental sequestration in a vaccine designed for women considering pregnancy.

Third, immunization studies show that vaccines already in hand can protect against malaria infection in animals models and in humans. Notably, consistent protection against experimental sporozoite challenge has been achieved in naive humans vaccinated with a recombinant anti-sporozoite vaccine based on the circumsporozoite protein — the principal surface antigen of sporozoites¹⁹. Protection has been achieved routinely in new world monkeys (*Aotus* and *Saimiri*) against blood-stage challenge with *P. falciparum* by vaccination with the antigen merozoite surface protein 1 (MSP1)²⁰. Both types of vaccine, anti-sporozoite and anti-blood-stage, have provided evidence of protection in field trials (see below), although the efficacy of these vaccines is still too low and the duration of protection still too short to be of practical value.

Last, studies have shown success in protecting mosquitoes against infection by *P. falciparum* and *P. vivax*. In these studies, the gametocyte-containing blood that is used as the infection source for the mosquito is mixed with serum from animals immunized with prototype human anti-mosquito-stage (transmission-blocking) vaccines^{5,21}. Thus, the efficacy of all three types of vaccine against human parasites — aimed at the sporozoite stage, blood stage and mosquito stage — has been proved in principle.

Strategies for vaccine design

Given that there is evidence supporting the feasibility of all three types of vaccine, what is the best strategy for designing them? What antigens should be included, and how should they be formulated? Certain principles can be defined that guide development.

First, we should choose antigens that meet specific criteria. Antigens are favoured as candidates if they are accessible to the immune system, induce protective immune responses in animal models, and either lack antigenic diversity or have at least limited diversity. Antigens such as PfEMP1, which show considerable

diversity but are potentially outstanding vaccine targets, will require a means by which to overcome that diversity — for example, by focusing on conserved functional domains²². Parasite receptors that mediate the binding of merozoites to erythrocytes, or of infected cells to endothelial cells, are important candidates. Some of these interactions are regulated by redundant pathways that may need to be blocked simultaneously (see review by Miller *et al.*, pages 673–679). Consideration of the parasite biology may lead to new antigens, such as proteins that are expressed in the mid-gut of mosquitoes and have therefore not evolved under human immune selection pressure, but which could be the targets of antibodies taken up in the mosquito's blood meal. Current candidate antigens are described in Table 1.

Second, we should combine antigens from different stages. Because of the stage specificity of antigen expression, the escape of even small numbers of parasites into the blood of a non-immune individual after immunization with a pre-erythrocytic-stage vaccine will result in clinical malaria. In recent experiments, naive volunteers challenged with 100–1,000 ring-infected erythrocytes (roughly 30–300-fold fewer than released from a single sporozoite-infected hepatocyte) had unaltered growth rates of the parasite during the pre-patent period and also developed clinical malaria at the expected threshold density of parasitaemia²³. But if a pre-erythrocytic vaccine were combined with a second line of defence — namely, a component designed to limit the growth of asexual blood stages — then vaccinated individuals with break-through blood-stage infection would be able to reduce further the risk of significant clinical illness. As another example, a combination of an anti-mosquito-stage vaccine with a blood- or pre-erythrocytic-stage vaccine would decrease the probability of resistant clones that emerge in vaccinated hosts spreading through the population.

Third, we should combine several antigens from a single stage. There are several arguments in favour of multivalency. (1) If the protection induced by each antigen is independent of that afforded by its companions, and if each is unable to induce sterile immunity on its own, then the potency of the vaccine will increase with more antigens. Few data support the supposition of independence, however, and problems with interference among antigens, if encountered, would have to be offset by the benefits of including several antigens or

Table 1 Candidate vaccine antigens

Stage and process	Antigens*	Primary mechanism
Sporozoite		
Hepatocyte invasion	CSP ⁷³ , TRAP/SSP2 ⁷³ , STARP ⁷⁴ , SALSA ⁷⁴	Antibody binding to antigens expressed on surface of sporozoite: Prevent binding interactions required for invasion Enhance splenic clearance or complement mediated lysis
Hepatic stages		
Growth, schizogony	CSP ⁷³ , TRAP/SSP2 ⁷³ , LSA1 ⁷³ , EXP1 ⁷³ , LSA3 ⁷⁵ , STARP ⁷⁴ , SALSA ⁷⁴	T-cell recognition of antigens expressed on surface of infected hepatocyte ⁷³ : Release soluble immune mediators (IFN- γ) resulting in intracellular parasite death Lyse infected hepatocyte directly (CTL)
Merozoite		
Erythrocyte invasion	MSP1 ²⁰ , MSP2 ³⁰ , MSP3 ⁷⁶ , MSP4 ⁷⁷ , PfEBA175 ⁶⁷ , PvDABP ⁷⁸ , AMA1 ⁷⁹ , SERA ⁸⁰ , GLURP ²⁸ , Pf155/RESA ³⁰ , RAP1 ⁸¹ , RAP2 ⁸¹	Antibody binding to parasite antigens: Agglutinate or prevent release of merozoites ⁸² Block invasion into erythrocytes ^{22,83,84} Induce monocytes to release soluble immune mediators, killing parasites (antibody-dependent cellular inhibition) ^{18,23} Facilitate phagocytosis ²³
Erythrocyte stages (asexual)		
Growth, sequestration	PfEMP1 ^{6,85} , rifins ²⁴ , Pf332 ²⁴	Antibody binding to antigens expressed on surface of infected erythrocyte: Enhance splenic clearance or complement mediated lysis Interfere with parasite nutrition and growth Induce antibody-dependent cellular inhibition Prevent or reverse endothelial adherence, thereby enhancing splenic clearance and reversing a key pathogenic mechanism
Erythrocyte rupture	Glycosylphosphatidylinositol (GPI) anchors ⁸⁶	Antibody binding to GPI resulting in prevention of toxic effects ⁸⁶
Mosquito stages		
Fertilization, oogenesis	Pfs25 ^{5,86} , Pfs28 ^{5,26} , Pfs48/45 ^{5,26} , Pfs230 ^{5,26}	Antibody binding to parasite antigens: Inhibit exflagellation and fertilization Complement induced lysis of gametes and zygotes Neutralize ookinete function

*Antigen names in bold type are in clinical development. Antigens have been listed according to the broad category of vaccine in which they are most likely to be included. However, many antigens are expressed in several stages (for example, LSA3 is expressed in blood stages as well as the liver, and most merozoite antigens are probably expressed in liver stages). Thus antigens in one category may also be targets in other stages.

by alterations to vaccine design that eliminate the interference. (2) Many leading vaccine candidates (MSP1 and apical membrane antigen 1 (AMA1)) are polymorphic, and several forms of each antigen may be required to provide protection against parasite diversity²⁴. (3) The immune response developed by each person is unique (owing to polymorphisms in HLA, other genetic traits and previous exposure); therefore mixtures of antigens may be required to ensure that each person has a reasonable chance of generating a sufficient response, particularly for genetically restricted T-cell responses. (4) Combining antigens may facilitate the development of a single vaccine that protects against more than one species of malaria. (5) The emergence of parasites that are simultaneously resistant to all immune responses induced by a multivalent vaccine is less likely than the emergence of parasites that are resistant to a single component, which could potentially prolong the useful life of the vaccine relative to that of a monovalent formulation.

Fourth, we need to keep vaccines as simple as possible. Optimizing the immunogenicity of individual components in a mixture is likely to be difficult, particularly as the complexity increases. For protein-based vaccines, the total amount of protein that can be injected in one vaccine is limited by the formulation and possibly by the frequency of adverse events; in addition, malaria vaccines will have to be inexpensive and the cost rises quickly with complexity. Vaccines using naked DNA or delivery systems such as replicons offer an attractive way of increasing the antigenic diversity within these constraints.

Last, we must induce the right kind of immune response. Biological characteristics of each parasite stage determine which immune responses are optimal for parasite destruction. For example, CD8-expressing T-cell and/or CD4-expressing T-cell type 1 responses may be essential for eliminating intracellular liver-stage parasites²⁵; by contrast, transmission-blocking vaccines must induce high, sustained amounts of antibody to neutralize antigens in the mosquito gut^{5,26}. Blood-stage vaccines also must generate high concentrations of antibody for protection, but may also need to induce other effector mechanisms, such as antibody-dependent cellular inhibition^{27,28}. Particular adjuvant systems may be needed to drive responses in one direction or the other, and this might constrain the composition of a particular vaccine.

So far, antigen choice has been dominated by the arbitrary order in which antigens have been identified. This is now changing with the sequencing of parasite genomes and an increasing understanding of parasite biology. We view the sequencing of the *P. falciparum* genome as a great boon to vaccine developers, because it should allow selection of the very best candidate antigens. Methods for screening the hundreds of proteins expressed during a particular parasite stage for those that induce protective immune responses are already being developed (see review by Hoffman *et al.*, pages 702–709).

These 'guiding principles' clearly involve compromises between conflicting demands. For example, the need to keep vaccines as simple as possible must be balanced against the need to cover antigenic diversity. Many researchers are proposing that a complex vaccine will be required, because this is the 'safe' assumption. Should one or a few antigens prove to elicit a universal protective response, vaccine development would be accelerated considerably.

Clinical trials of malaria vaccines

Over the past 15 years, a series of phase 1/2 vaccine trials has been reported using synthetic peptides or recombinant proteins based on malarial antigens. Engers and Godal²⁹ report that there have been roughly 40 trials up to 1998. Most trials were directed against sporozoites or liver stages, in which the use of experimental mosquito challenges allows rapid progress through phase 1 to phase 2a preliminary efficacy studies. Anti-sporozoite vaccines tested have included completely synthetic peptides, conjugates of synthetic peptide with proteins such as tetanus toxoid to provide T-cell help, recombinant malaria proteins, particle-forming recombinant chimaeric constructs, recombinant viruses, and bacteria- and DNA-

based vaccines. Several trials of asexual blood-stage vaccines have used either synthetic peptide conjugates or recombinant proteins, and there has been one trial of a transmission-blocking vaccine — recombinant Pfs25 (ref. 26; and unpublished data from National Institutes of Health (NIH) Pfs25 clinical trial). A recurring theme is the difficulty of obtaining a sufficiently strong and long-lasting immune response in humans even though the same vaccine preparation is often strongly immunogenic in test animals³⁰. Various strategies have been designed to overcome this limitation, including the exploration of potent immune-stimulatory conjugates or adjuvants to boost the human response. An interesting approach is the *in situ* expression of the vaccine through attenuated strains of *Salmonella*³¹.

Vaccines directed against the circumsporozoite protein covering the sporozoite illustrate the frustrating quest for a consistently protective immune response in humans. Early studies with both recombinant proteins, peptide conjugates and recombinant protein conjugates could elicit anti-circumsporozoite antibody, provided marginal protection in phase 2a studies but provided no protection in field studies³². More recently, collaboration between Glaxo-SmithKline and the US Army has produced a chimaeric protein, consisting of a fusion between the circumsporozoite protein and the hepatitis B surface antigen (HBsAg), which is expressed in yeast¹⁹. In the presence of unmodified HBsAg, the mixture forms typical HBsAg particles (the RTS,S vaccine). Although much more immunogenic than recombinant circumsporozoite protein, RTS,S is still incapable of eliciting a protective response when used with conventional adjuvants, but elicits a protective response in about half of vaccine recipients with the AS02 (formerly SBAS2) experimental oil in water adjuvant, which also contains monophosphoryl lipid A (3D-MPL) and the saponin QS21. Other recent approaches to a circumsporozoite vaccine have used synthetic peptide technology to generate immunogens containing both T- and B-cell epitopes with an endogenous lipophilic adjuvant³³.

Only a few *P. falciparum* vaccines have been tested in field trials. This area has been dominated by the synthetic peptide vaccine SPf66, which is absorbed on alum and is directed at blood-stage parasites³⁴. Initial results of trials in children and adults residing in relatively low endemic regions of South America and children in a highly endemic area of Tanzania suggested that this vaccine might delay the time to a first clinical malaria episode, but subsequent trials in infants in The Gambia and Tanzania, children in Thailand and adults in Brazil failed to show efficacy. In a recent meta-analysis of nine trials involving 9,800 volunteers, Graves and Gelbrand³² concluded that there was no evidence of efficacy for SPf66 in Africa but that it caused a modest reduction of malaria attacks in South America.

Two other vaccines that have recently undergone field trials show promise. In a collaboration with the Medical Research Council (MRC) unit in The Gambia, a trial of the RTS,S vaccine in adult Gambian males resulted in a significant reduction in the rate at which they were infected after vaccination, although the protection was not long-lived³⁵. A combination of three asexual blood-stage antigens (the 190L fragment of MSP1, one form of the polymorphic MSP2 antigen and a portion of the RESA or ring-infected erythrocyte surface antigen)³⁰ has been tested in children aged 5–9 years in the Wosera district of Papua New Guinea. Designed to determine the parasitological outcome, this trial detected significant decreases in the parasite density and the frequency of parasite episodes greater than 1,000 per microlitre, and a major switch in the MSP2 genotype of parasites³⁶.

Although protein-based vaccines are promising and are furthest along the pathway towards malaria vaccine development, alternative technologies such as DNA-based vaccines may prove particularly amenable to multivalent formulations and ultimately less expensive to produce, store and deliver than conventional vaccines (ref. 37; and Box 2). Progress with DNA vaccines has been rapid, and studies have demonstrated the following. (1) Protection against *P. yoelii*

sporozoite challenge in BALB/c mice immunized with a plasmid encoding the *P. yoelii* circumsporozoite protein in 1994 (ref. 38). (2) Improved protection in mice with a two-plasmid vaccine in 1996 (ref. 39). (3) Complete protective efficacy of DNA prime/poxvirus boost regimens in mice in 1998 (ref. 40). (4) Safety of DNA and induction of genetically restricted, antigen-specific, cytotoxic T lymphocyte (CTL) responses in humans vaccinated with a plasmid encoding *P. falciparum* circumsporozoite protein in 1998 (refs 41,42). (5) Induction of CD8-expressing and CD4-expressing T-cell-dependent interferon- γ responses by ELISPOT (enzyme-linked immunospot) assay in humans immunized with the same plasmid in 2001 (ref. 43). (6) Partial protection of rhesus monkeys against *P. knowlesi* sporozoite challenge with a four-plasmid cocktail boosted by poxvirus constructs encoding the priming malarial antigens in 2001 (ref. 44; and W. O. Rogers, unpublished data). (7) Similar success with *P. cynomolgi* challenge of rhesus monkeys immunized with a epitope string derived from *P. falciparum* fused with the cholera toxin B subunit in 2000 (ref. 45). A DNA vaccine including a string of CTL epitopes from six different *P. falciparum* pre-erythrocytic antigens fused to the full-sequence thrombospondin-related adhesion protein (TRAP) from *P. falciparum*, also administered using a plasmid-prime, poxvirus-boost approach, has been in clinical evaluation in Oxford since 1999 and in The Gambia since September 2000, with encouraging results⁴⁶.

DNA plasmids encoding malaria antigens have generated antibody responses in animal models^{38,47–50}, but not in humans as yet⁴¹. The interferon- γ responses shown in humans are appropriate for liver-stage vaccines designed to effect intracellular killing, but are probably insufficient for vaccines designed to neutralize circulating sporozoites or blood-stage parasites. Many new modalities to improve expression and immunogenicity and to extend the breadth of the immune response to DNA-based vaccines are now being tested. These include replacing native genes with synthetic genes using codon frequencies optimized for host expression⁵¹, adding sequences such as the Fc region of immunoglobulin that increase antigen uptake by dendritic cells⁵², absorbing plasmids onto microparticles⁵³ to increase antigen uptake by dendritic cells, particle-mediated gene transfer (gene gun)⁵⁴ or jet injection⁴³ to enhance plasmid delivery, and co-stimulation with plasmids encoding cytokines, such as granulocyte/macrophage colony-stimulating factor^{55–57}, or other immunostimulatory molecules such as CpG motifs⁵⁸ or an interleukin-2/immunoglobulin fusion protein⁵⁹.

Of the various approaches, the one that is most potent in improving protection in animal models is heterologous vaccination, in which DNA plasmids are used to prime specific immune responses and virus constructs^{44,50,54,60} or recombinant proteins⁶¹ are used to boost these responses. Several viral vectors are under development, including avian and mammalian recombinant poxviruses^{62,63}, non-replicating and replicating recombinant adenoviruses⁶⁴, and alphavirus replicons⁶⁵. Clinical studies of DNA priming followed by either recombinant poxvirus or protein boosting are in progress at various centres. Although DNA (plasmid) vaccines on their own are currently not sufficiently immunogenic or protective, if immunogenicity can be improved then DNA-based technologies should allow the development of effective malaria vaccines.

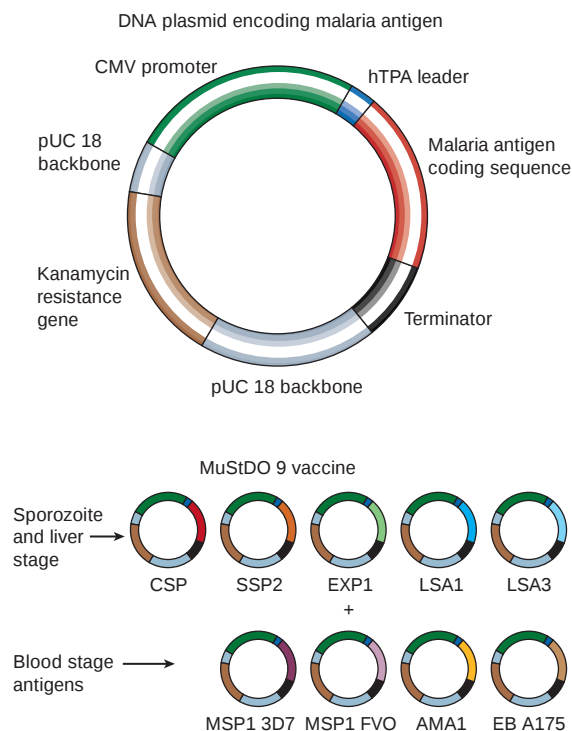
Challenges in vaccine development

Our progress in identifying antigens and the demonstration that several of these can elicit protective immune responses give us reason for optimism. Nevertheless, a substantial investment of resources is needed to improve the degree and duration of protection and to bring vaccines to licensure. But perceptions concerning the economics of developing, manufacturing and distributing vaccines to a global market that comprises some of the world's poorest people limit our ability to mobilize these resources, especially from the private sector.

Typically, pharmaceuticals have been developed through a process that involves screening large numbers of compounds for

Box 2

Cocktail to soothe tropical fevers?



Supercoiled rings of double-stranded DNA containing malaria genes are injected, taken up by muscle cells and translated into malarial proteins. These proteins are processed and presented by antigen-presenting cells, inducing immune responses designed to inhibit the developing malaria parasites. The MuStDO 9 vaccine^{87,88} consists of five plasmids encoding proteins from the sporozoite and liver stages of *P. falciparum*, CSP, SSP2, EXP1, LSA1 and LSA3, and four encoding proteins from blood stages, MSP1 (two alleles representing the major antigenic families), AMA1 and EBA175. The two-tiered design is based on the rationale that parasites escaping the pre-erythrocytic-stage defence will be suppressed by a second line of immune responses directed at blood stages, thus decreasing the likelihood of serious illness in vaccinated individuals with breakthrough blood-stage infections. Each gene in the expression cassette has been synthesized using codon frequencies that are consistent with mammalian patterns. At present, naked DNA vaccines are not sufficiently immunogenic to protect humans against malaria, and require boosting by recombinant viral constructs or recombinant proteins to achieve measurable efficacy. New formulations and designs aiming to increase plasmid uptake by dendritic cells, improve expression and strengthen antibody responses may remove this requirement. So far, DNA plasmids encoding malarial antigens have been administered to roughly 80 healthy adult volunteers in the United States by the US Navy and about 50 healthy adult volunteers in Britain or The Gambia (mainly in heterologous prime-boost regimens with modified vaccinia virus Ankara) by the Oxford University group (A. V. S. Hill, personal communication) and seem to be safe and well tolerated⁴¹.

activity and picking lead candidates for further development. For malaria we now have at least 5,000 candidate proteins and multiple variants of many of these, and although we have *in vitro* and animal models that give indications of a candidate's importance, we lack validated models that reliably predict what protection a vaccine formulation will give in humans. Views differ on the contribution

Box 3

Malaria vaccines in or close to clinical trials

1. Pre-erythrocytic vaccines RTS,S and TRAP/SSP2, and blood-stage vaccines MSP1₄₂-3D7 and AMA1-3D7 by a collaboration between the US Army Walter Reed Army Institute of Research (WRAIR), USAID and GlaxoSmithKline.
2. An RTS,S phase 2b trial in Gambian adults and a phase 1 trial in Gambian children being undertaken by the MRC unit in The Gambia, in collaboration with GlaxoSmithKline and with support from WRAIR, the University of Oxford, and The London School of Hygiene and Tropical Medicine.
3. Plasmid DNA, modified vaccinia virus Ankara and an attenuated fowlpox strain encoding the multi-epitope-TRAP insert, by a collaboration between the University of Oxford and the MRC, with phase 1/2a trials in Britain and phase 1 trials in The Gambia⁴⁶.
4. Phase 1/2a studies of multistage DNA vaccine combinations ('MuStDO') in the United States and Ghana by a collaboration between the US Navy and USAID^{87,88}.
5. Phase 1/2a studies of modified vaccinia virus Ankara and a fowlpox strain encoding the circumsporozoite protein by the University of Oxford.

6. Vaccines based on the merozoite surface protein MSP1₁₉ at the Institut Pasteur.
 7. Two phase 1 trials by the European Malaria Vaccine Initiative with protein vaccines made by peptide synthesis (MSP3, GLURP), followed by a phase 1 trial of Pichia-expressed recombinant AMA1.
 8. Phase 1 studies of *P. vivax* and *P. falciparum* transmission-blocking protein vaccines Pvs25 and Pfs25 by the NIH Malaria Vaccine Development Unit.
 9. Phase 1 studies of a pre-erythrocytic-stage vaccine that is based on a virus-like particle comprising a modified hepatitis B core particle and circumsporozoite protein-specific T- and B-cell epitopes by a collaboration between Apovia, New York University, Malaria Vaccine Initiative, NIH and the University of Maryland.
- Several other antigens are at an advanced state of development and many phase 1 and some phase 2 clinical trials could reasonably be expected in the next 1–2 years. These include major polymorphisms of AMA1, MSP1₄₂ and MSP2 that explore different expression systems, and single forms of other antigens such as MSP4, RAP2, EBA175 and the *P. vivax* Duffy-binding protein.

that can be made by new and old world primate models to screen candidate vaccines^{66,67}. Even where primate models assist antigen selection, the development of useful formulations and vaccine regimens will require several human phase 1 and phase 2 trials, and screening for protection in humans is difficult.

Screening anti-morbidity vaccines in human populations in endemic areas imposes many restraints. First, producing multiple test vaccines is a big task. Second, these vaccines are designed primarily for children, and using children for a primary efficacy screen contributes to the complexity of the trial design. Last, decisions need to be made regarding the end point of the trial. Both for ethical reasons and to diminish sample size requirements, surrogate markers of efficacy may be chosen.

For example, in the Papua New Guinea trial discussed above³⁶, parasite density was used as the test of efficacy. This required group sizes of 30 with a 12-week follow up to give reasonable power of detecting a 30% decrease in parasite density. In the same population⁶⁸, groups of a few hundred would have been required to see a significant decrease in febrile cases if the vaccine caused a 30% decrease in malaria-related fever. In this region, the frequency of severe disease or death from malaria is too low to use as a useful end point, even if this could have been justified ethically. In much of Africa — which has an infant mortality of about 100 in 1,000 live births⁶⁹ — depending on the accuracy with which a cause of death can be diagnosed, group sizes of many thousands to tens of thousands would be needed to use mortality as an end point, and this is not feasible at an early stage of vaccine development. Because there are gaps in our understanding of the progression of pathology from parasitaemia to death (see review by Miller *et al.*, pages 673–679), our choice of end point for early efficacy studies is associated with a risk of either discarding a good vaccine because it fails to give an imperfect correlate of protection in early stage testing or wasting scarce resources by taking a poor vaccine through extensive clinical testing.

In contrast to anti-morbidity vaccines for infants or other high risk groups, anti-infection vaccines designed for malaria-naïve travellers and residents of low endemic regions face more stringent efficacy constraints, but should be relatively simpler to test. Experimental sporozoite challenge of vaccinated, naïve volunteers has become a standard way of early stage testing of vaccines for travellers, especially pre-erythrocytic-stage vaccines⁷⁰, and a blood-stage challenge has also been developed as an alternative for testing blood-stage vaccines²³.

In principle, mosquito-stage vaccines should have the simplest development path, because there is a relatively straightforward test for assessing the impact of vaccine-induced antibodies on the transmission of parasites to mosquitoes, allowing down-selection to take place at an earlier stage of development²⁶. The eventual proof that such a vaccine is effective under field conditions will require large trials in which a high proportion of the population is vaccinated and whole villages are randomized to treatment groups.

Meeting the challenge

The development of vaccines for the billion people living in endemic countries will require a different scheme from that of more commercially attractive vaccines. The public sector will need to provide the 'push' to facilitate vaccine technology development and early-stage testing and also the 'pull' of guaranteed markets to encourage pharmaceutical firms to license and manufacture vaccines⁵.

The manufacturing capacity to meet global demands may be relatively modest, depending on the type of vaccine. Most of the world's need for a vaccine to prevent serious disease and death in infants will be in sub-Saharan Africa where the live birth rate is 26 million per year. For a recombinant protein vaccine, roughly 1 kg of each protein component would be needed each year, depending on the number and size of the doses. Even with the non-optimized expression systems that are being used to produce prototype vaccines, this could be met by a single plant capable of fermenting 200 litres per week, which is well within the capacity of biotechnology companies in more industrialized malaria-endemic countries in which national considerations, economics and marketing strategies make local production attractive.

Currently, the biggest bottleneck is the capacity for producing clinical grade material for protein-based vaccines and for performing field trials for all types of vaccine. One of the most encouraging recent developments has been the substantial expansion in government and philanthropic interests to address this need, including the founding of the European Malaria Vaccine Initiative⁷¹ and the creation of the Malaria Vaccine Initiative at the Program for Appropriate Technology in Health through a grant from the Bill & Melinda Gates Foundation⁷². The US NIH has developed new programmes for the development and testing of malaria vaccines through its extramural programme and through the creation of an intramural Malaria Vaccine Development Unit. These new developments join several continuing programmes, such as those of the World Health

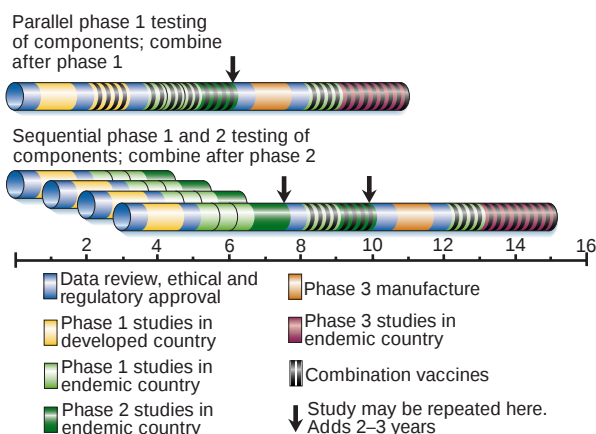


Figure 2 Timetable of clinical trials. Shown is an optimistic and simplified view of the timetable to completion of phase 3 (pre-registration) trials for a hypothetical tetravalent malaria vaccine starting from the time that clinical grade material is available. Each stage in the process requires review and approval by regulatory bodies. Initial phase 1 safety, immunogenicity and dose escalation testing in adults in the country of manufacture will generally need to be repeated in endemic countries in adults, children and infants, and these trials must run sequentially for safety reasons. The vaccine is then tested for efficacy, and if successful will move to larger phase 3 studies before registration, to confirm its efficacy and safety. Phase 3 trials test the vaccine made under the same conditions that will be used for final manufacture; before they can commence, the final manufacturing process and plant need to be established. Two schemes are examined here. The first combines the four components of the vaccine after the initial phase 1 trial (combination vaccine shown with banding). The second combines the vaccine only after each component is shown to be efficacious. Parallel phase 1 testing takes less time than sequential testing, but requires more resources. Early combination takes fewer resources in field studies, but may require separate studies to show that all components are required (not shown on time line).

Organization, US Agency for International Development (USAID) and the US Department of Defense. Several sites in endemic areas such as The Gambia, Ghana, Kenya, Mali, Mozambique, Papua New Guinea and Tanzania are suitable for vaccine testing.

A potential capacity for vaccine testing has been created through the development of the African Malaria Vaccine Testing Network and its programme of training Africans in clinical trial capability. In addition, many endemic countries are making significant contributions from their scarce resources through the support of staff and facilities for malaria research in general and for vaccines in particular. Despite these developments, progress remains limited by resources, and this limitation will become more acute as vaccine testing programmes gain momentum. Put into perspective, the public support for development of malaria vaccines is roughly tenfold less than the support for a vaccine for HIV — a disease of comparable global impact.

In this setting, several groups have phase 1/2 testing programmes either underway, or have clinical grade immunogens prepared with phase 1 trials commencing in the near future. These developments are summarized in Box 3.

When will there be malaria vaccines?

On the assumption that at least some of these vaccines will be efficacious, when can we expect the vaccine? Even under the most optimistic scheme of unlimited resources (Fig. 2), it will still be many years from now, requiring iterative testing of improved combinations and formulations until sufficient efficacy is obtained. In the meantime there is much to do to ensure that when the vaccines are available, health infrastructures are in place to deliver integrated

programmes that will use the vaccines effectively. That may be as much of a challenge as the vaccine itself.

- Jeffery, G. M. Epidemiological significance of repeated infections with homologous and heterologous strains and species of *Plasmodium*. *Bull. World Health Organ.* **35**, 873–882 (1966).
- Clyde, D. F. Immunity to falciparum and vivax malaria induced by irradiated sporozoites: a review of the University of Maryland studies, 1971–75. *Bull. World Health Organ.* **68**(Suppl.), 9–12 (1990).
- Marsh, K. & Snow, R. W. Malaria transmission and morbidity. *Parasitologia* **41**, 241–246 (1999).
- Ricke, C. H. *et al.* Plasma antibodies from malaria-exposed pregnant women recognize variant surface antigens on *Plasmodium falciparum*-infected erythrocytes in a parity-dependent manner and block parasite adhesion to chondroitin sulfate A. *J. Immunol.* **165**, 3309–3316 (2000).
- Carter, R., Mendis, K. N., Miller, L. H., Molineaux, L. & Saul, A. Malaria transmission-blocking vaccines—how can their development be supported? *Nature Med.* **6**, 241–244 (2000).
- Saul, A. The role of variant surface antigens on malaria-infected red blood cells. *Parasitol. Today* **15**, 455–457 (1999).
- Miller, L. H. Evolution of the human genome under selective pressure from malaria: applications for control. *Parasitologia* **41**, 77–82 (1999).
- Nussenzeig, R. S., Vanderberg, J., Most, H. & Orton, C. Protective immunity produced by the injection of X-irradiated sporozoites of *Plasmodium berghei*. *Nature* **216**, 160–162 (1967).
- Collins, W. E. & Contacos, P. G. Immunization of monkeys against *Plasmodium cynomolgi* by X-irradiated sporozoites. *Nature* **236**, 176–177 (1972).
- Egan, J. E. *et al.* Humoral immune responses in volunteers immunized with irradiated *Plasmodium falciparum* sporozoites. *Am. J. Trop. Med. Hyg.* **49**, 166–173 (1993).
- Rieckmann, K. H., Beaudoin, R. L., Cassels, J. S. & Sell, K. W. Use of attenuated sporozoites in the immunization of human volunteers against falciparum malaria. *Bull. World Health Organ.* **57**(Suppl. 1), 261–265 (1979).
- Hoffman, S. L. *et al.* Sporozoite vaccine induces genetically restricted T cell elimination of malaria from hepatocytes. *Science* **244**, 1078–1081 (1989).
- Hoffman, S. L. *et al.* Protection of humans against malaria by immunization with radiation-attenuated *Plasmodium* sp. sporozoites. *J. Infect. Dis.* (in the press).
- Baird, J. K. Host age as a determinant of naturally acquired immunity to *Plasmodium falciparum*. *Parasitol. Today* **11**, 105–111 (1995).
- Hoffman, S. L. *et al.* Naturally acquired antibodies to sporozoites do not prevent malaria: vaccine development implications. *Science* **237**, 639–642 (1987).
- Cohen, S., McGregor, I. A. & Carrington, S. Gamma globulin and acquired immunity to human malaria. *Nature* **192**, 733–737 (1961).
- Edozien, J. C., Gilles, H. M. & Udeozo, I. O. K. Adult and cord-blood gamma globulin and immunity to malaria in Nigerians. *Lancet* **ii**, 951–955 (1962).
- Sabchareon, A. *et al.* Parasitologic and clinical human response to immunoglobulin administration in falciparum malaria. *Am. J. Trop. Med. Hyg.* **45**, 297–308 (1991).
- Kester, K. E. *et al.* Efficacy of recombinant circumsporozoite protein vaccine regimens against experimental *Plasmodium falciparum* malaria. *J. Infect. Dis.* **183**, 640–647 (2001).
- Stowers, A. W. *et al.* Efficacy of two alternate vaccines based on *Plasmodium falciparum* merozoite surface protein 1 in an *Aotus* challenge trial. *Infect. Immun.* **69**, 1536–1546 (2001).
- Hisaeda, H. *et al.* Antibodies to malaria vaccine candidates Pvs25 and Pvs28 completely block the ability of *Plasmodium vivax* to infect mosquitoes. *Infect. Immun.* **68**, 6618–6623 (2000).
- Gamain, B., Miller, L. H. & Baruch, D. I. The surface variant antigens of *Plasmodium falciparum* contain cross-reactive epitopes. *Proc. Natl Acad. Sci. USA* **98**, 2664–2669 (2001).
- Cheng, Q. *et al.* Measurement of *Plasmodium falciparum* growth rate in vivo: a test of malaria vaccines. *Am. J. Trop. Med. Hyg.* **57**, 495–500 (1997).
- Bolad, A. & Berzins, K. Antigenic diversity of *Plasmodium falciparum* and antibody-mediated parasite neutralization. *Scand. J. Immunol.* **52**, 233–239 (2000).
- Doolan, D. L. & Hoffman, S. L. The complexity of protective immunity against liver-stage malaria. *J. Immunol.* **165**, 1453–1462 (2000).
- Carter, R. Transmission blocking malaria vaccines. *Vaccine* **19**, 2309–2314 (2001).
- Badell, E. *et al.* Human malaria in immunocompromised mice: an *in vivo* model to study defense mechanisms against *Plasmodium falciparum*. *J. Exp. Med.* **192**, 1653–1660 (2000).
- Oeuvray, C. *et al.* Cytophilic immunoglobulin responses to *Plasmodium falciparum* glutamate-rich protein are correlated with protection against clinical malaria in Dielmo, Senegal. *Infect. Immun.* **68**, 2617–2620 (2000).
- Engers, H. D. & Godal, T. Malaria vaccine development: current status. *Parasitol. Today* **14**, 56–64 (1998).
- Saul, A. *et al.* Human phase I vaccine trials of 3 recombinant asexual stage malaria antigens with Montanide ISA720 adjuvant. *Vaccine* **17**, 3145–3159 (1999).
- Wu, S. *et al.* Construction and immunogenicity in mice of attenuated *Salmonella typhi* expressing *Plasmodium falciparum* merozoite surface protein 1 (MSP-1) fused to tetanus toxin fragment C. *J. Biotechnol.* **83**, 125–135 (2000).
- Graves, P. & Gelbrand, H. Vaccines for preventing malaria. Cochrane Review, The Cochrane Library, issue no. 3 <<http://update-software.com/cochrane/abstracts/ab000129.htm>> (Update Software, Oxford, 2001).
- Nardin, E. H. *et al.* A totally synthetic polyoxime malaria vaccine containing *Plasmodium falciparum* B cell and universal T cell epitopes elicits immune responses in volunteers of diverse HLA types. *J. Immunol.* **166**, 481–489 (2001).
- Patarroyo, M. E. *et al.* A synthetic vaccine protects humans against challenge with asexual blood stages of *Plasmodium falciparum* malaria. *Nature* **332**, 158–161 (1988).
- Bojang, K. A. *et al.* Efficacy of RTS,S/AS02 malaria vaccine against *P. falciparum* infection in semi-immune adult men in The Gambia: a randomised trial. *Lancet* **358**, 1927–1934 (2001).
- Genton, B. *et al.* A recombinant blood-stage malaria vaccine reduces *Plasmodium falciparum* density and exerts selective pressure on parasite populations in a Phase I/IIb trial in Papua New Guinea. *J. Infect. Dis.* (in the press).
- Doolan, D. L. *et al.* DNA vaccination as an approach to malaria control: current status and strategies. *Curr. Top. Microbiol. Immunol.* **226**, 37–56 (1998).
- Seidegh M., Hedstrom R., Hobart P. & Hoffman S.L. Protection against malaria by immunization with plasmid DNA encoding circumsporozoite protein. *Proc. Natl Acad. Sci. USA* **91**, 9866–9870 (1994).

39. Doolan, D. L. *et al.* Circumventing genetic restriction of protection against malaria with multigene DNA immunization: CD8⁺ cell-, interferon γ -, and nitric oxide-dependent immunity. *J. Exp. Med.* **183**, 1739–1746 (1996).
40. Schneider, J. *et al.* Enhanced immunogenicity for CD8⁺ T cell induction and complete protective efficacy of malaria DNA vaccination by boosting with modified vaccinia virus Ankara. *Nature Med.* **4**, 397–402 (1998).
41. Le, T. P. *et al.* Safety, tolerability and humoral immune responses after intramuscular administration of a malaria DNA vaccine to healthy adult volunteers. *Vaccine* **18**, 1893–1901 (2000).
42. Wang, R. *et al.* Induction of antigen-specific cytotoxic T lymphocytes in humans by a malaria DNA vaccine. *Science* **282**, 476–480 (1998).
43. Wang, R. *et al.* Induction of CD4⁺ T cell dependent CD8⁺ type 1 responses in humans by a malaria DNA vaccine. *Proc. Natl Acad. Sci. USA* **98**, 10817–10822 (2001).
44. Rogers, W. O. *et al.* Multi-stage multi-antigen heterologous prime boost vaccine for *Plasmodium knowlesi* malaria provides partial protection in rhesus macaques. *Infect. Immun.* **69**, 5565–5572 (2001).
45. Zhong, H. *et al.* Induction of protective immune responses in rhesus monkey by immunization with recombinant plasmids of polyvalent epitopes of *Plasmodium falciparum* using cholera toxin B as adjuvant. [Translation] *Yi Chuan Xue Bao [Acta Genetica Sinica]* **27**, 966–971 (2000).
46. Ferry, G. First DNA malaria vaccine on trial in Africa. *Curr. Biol.* **10**, R810–R811 (2000).
47. Gramzinski, R. A. *et al.* Malaria DNA vaccines in *Aotus* monkeys. *Vaccine* **15**, 913–915 (1997).
48. Shi, Y. P. *et al.* Immunogenicity and *in vitro* protective efficacy of a recombinant multistage *Plasmodium falciparum* candidate vaccine. *Proc. Natl Acad. Sci. USA* **96**, 1615–1620 (1999).
49. Aguiar, J. C. *et al.* Enhancement of the immune response in rabbits to a malaria DNA vaccine by immunization with a needle-free jet device. *Vaccine* **20**, 275–280 (2001).
50. Schneider, J. *et al.* A prime-boost immunisation regimen using DNA followed by recombinant modified vaccinia virus Ankara induces strong cellular immune responses against the *Plasmodium falciparum* TRAP antigen in chimpanzees. *Vaccine* **19**, 4595–4602 (2001).
51. Narum, D. L. *et al.* Codon optimization of gene fragments encoding *Plasmodium falciparum* merozoite proteins enhances DNA vaccine protein expression and immunogenicity in mice. *Infect. Immun.* **69**, 7250–7253 (2001).
52. You, Z., Huang, X., Hester, J., Toh, H. C. & Chen, S. Y. Targeting dendritic cells to enhance DNA vaccine potency. *Cancer Res.* **61**, 3704–3711 (2001).
53. O'Hagan, D. *et al.* Induction of potent immune responses by cationic microparticles with adsorbed human immunodeficiency virus DNA vaccines. *J. Virol.* **75**, 9037–9043 (2001).
54. Degano, P., Schneider, J., Hannan, C. M., Gilbert, S. C. & Hill, A. V. Gene gun intradermal DNA immunization followed by boosting with modified vaccinia virus Ankara: enhanced CD8⁺ T cell immunogenicity and protective efficacy in the influenza and malaria models. *Vaccine* **18**, 623–632 (1999).
55. Sakai, T. *et al.* DNA immunization with *Plasmodium falciparum* serine repeat antigen: regulation of humoral immune response by coinoculation of cytokine expression plasmid. *Parasitol. Int.* **48**, 27–33 (1999).
56. Weiss, W. R. *et al.* A plasmid encoding murine granulocyte-macrophage colony-stimulating factor increases protection conferred by a malaria DNA vaccine. *J. Immunol.* **161**, 2325–2332 (1998).
57. Sedegah, M. *et al.* Improving protective immunity induced by DNA-based immunization: priming with antigen and GM-CSF-encoding plasmid DNA and boosting with antigen-expressing recombinant poxvirus. *J. Immunol.* **164**, 5905–5912 (2000).
58. Jones, T. R. *et al.* Synthetic oligodeoxynucleotides containing CpG motifs enhance immunogenicity of a peptide malaria vaccine in *Aotus* monkeys. *Vaccine* **17**, 3065–3071 (1999).
59. Barouch, D. H. *et al.* Control of viremia and prevention of clinical AIDS in rhesus monkeys by cytokine-augmented DNA vaccination. *Science* **290**, 486–492 (2000).
60. Schneider, J. *et al.* Induction of CD8⁺ T cells using heterologous prime-boost immunisation strategies. *Immunol. Rev.* **170**, 29–38 (1999).
61. Jones, T. R. *et al.* Protection of *Aotus* monkeys by *Plasmodium falciparum* EBA-175 region II DNA prime-protein boost immunization regimen. *J. Infect. Dis.* **183**, 303–312 (2001).
62. Dong, W. *et al.* Assessment of a vaccinia virus vectored multi-epitope live vaccine candidate for *Plasmodium falciparum*. *Int. J. Parasitol.* **31**, 57–62 (2001).
63. Ockenhouse, C. F. *et al.* Phase I/IIa safety, immunogenicity, and efficacy trial of NYVAC-Pf7, a pox-vectored, multiantigen, multistage vaccine candidate for *Plasmodium falciparum* malaria. *J. Infect. Dis.* **177**, 1664–1673 (1998).
64. Rodrigues, E. G. *et al.* Interferon- γ -independent CD8⁺ T cell-mediated protective anti-malaria immunity elicited by recombinant adenovirus. *Parasite Immunol.* **22**, 157–160 (2000).
65. Davis, N. L. *et al.* Vaccination of macaques against pathogenic simian immunodeficiency virus with Venezuelan equine encephalitis virus replicon particles. *J. Virol.* **74**, 371–378 (2000).
66. Stowers, A. W. & Miller, L. H. Are trials in new world monkeys on the critical path for blood-stage malaria-vaccine development? *Trends Parasitol.* **17**, 415–419 (2001).
67. Heppner, D. G. *et al.* New world monkey efficacy trials for malaria vaccine development: critical path or detour? *Trends Parasitol.* **17**, 419–425 (2001).
68. Genton, B. *et al.* The epidemiology of malaria in the Wosera area, East Sepik Province, Papua New Guinea, in preparation for vaccine trials. II. Mortality and morbidity. *Ann. Trop. Med. Parasitol.* **89**, 377–390 (1995).
69. Smith, T. A., Leuenberger, R. & Lengeler, C. Child mortality and malaria transmission intensity in Africa. *Trends Parasitol.* **17**, 145–149 (2001).
70. Church, L. W. *et al.* Clinical manifestations of *Plasmodium falciparum* malaria experimentally induced by mosquito challenge. *J. Infect. Dis.* **175**, 915–920 (1997).
71. Hagan, P., Bjorvatn, B. & Jepsen, S. European malaria vaccine initiative. *Parasitol. Today* **15**, 47–48 (1999).
72. Nossal, G. J. The global alliance for vaccines and immunization—a millennial challenge. *Nature Immunol.* **1**, 5–8 (2000).
73. Hoffman, S. L., Franke, E. D., Hollingdale, M. R. & Druilhe, P. in *Malaria Vaccine Development: a Multi-immune Response Approach* (ed. Hoffman, S. L.) 35–75 (ASM, Washington, 1996).
74. Perlaza, B. L. *et al.* Immunogenicity of four *Plasmodium falciparum* pre-erythrocytic antigens in *Aotus lemurinus* monkeys. *Infect. Immun.* **66**, 3423–3428 (1998).
75. Daubersies, P. *et al.* Protection against *Plasmodium falciparum* malaria in chimpanzees by immunization with the conserved pre-erythrocytic liver-stage antigen 3. *Nature Med.* **6**, 1258–1263 (2000).
76. Oeuvray, C. *et al.* Merozoite surface protein-3: a malaria protein inducing antibodies that promote *Plasmodium falciparum* killing by cooperation with blood monocytes. *Blood* **84**, 1594–1602 (1994).
77. Wang, L., Richie, T. L., Stowers, A., Nhan, D. H. & Coppel, R. L. Naturally acquired antibody responses to *Plasmodium falciparum* merozoite surface protein 4 in a population living in an area of endemicity in Vietnam. *Infect. Immun.* **69**, 4390–4397 (2001).
78. Singh, S. *et al.* Biochemical, biophysical and functional characterization of bacterially expressed and refolded binding domain of *Plasmodium vivax* Duffy binding protein. *J. Biol. Chem.* **276**, 17111–17116 (2001).
79. Hodder, A. N., Crewther, P. E. & Anders, R. F. Specificity of the protective antibody response to apical membrane antigen 1. *Infect. Immun.* **69**, 3286–3294 (2001).
80. Fox, B. A., Xing, L. P., Suzue, K., Horii, T. & Bzik, D. J. *Plasmodium falciparum*: an epitope within a highly conserved region of the 47-kDa amino-terminal domain of the serine repeat antigen is a target of parasite-inhibitory antibodies. *Exp. Parasitol.* **85**, 121–134 (1997).
81. Collins, W. E. *et al.* Efficacy of vaccines containing rhoptry-associated proteins RAP1 and RAP2 of *Plasmodium falciparum* in *Saimiri boliviensis* monkeys. *Am. J. Trop. Med. Hyg.* **62**, 466–479 (2000).
82. Lyon, J. A., Thomas, A. W., Hall, T. & Chulay, J. D. Specificities of antibodies that inhibit merozoite dispersal from malaria-infected erythrocytes. *Mol. Biochem. Parasitol.* **36**, 77–85 (1989).
83. Saul, A. & Miller, L. H. A Robust neutralization test for *Plasmodium falciparum* malaria. *J. Exp. Med.* **193**, F51–F54 (2001).
84. O'Donnell, R. A. *et al.* Antibodies against MSP-1₁₉ are a major component of the invasion-inhibitory response in individuals immune to malaria. *J. Exp. Med.* **193**, 1403–1412 (2001).
85. Scherf, A., Pouvelle, B., Buffet, P. A. & Gysin, J. Molecular mechanisms of *Plasmodium falciparum* placental adhesion. *Cell Microbiol.* **3**, 125–131 (2001).
86. Schofield, L. *et al.* Neutralizing monoclonal antibodies to glycosylphosphatidylinositol, the dominant TNF- α -inducing toxin of *Plasmodium falciparum*: prospects for the immunotherapy of severe malaria. *Ann. Trop. Med. Parasitol.* **87**, 617–626 (1993).
87. Doolan, D. L. & Hoffman, S. L. DNA-based vaccines against malaria: status and promise of the multi-stage malaria DNA vaccine operation. *Int. J. Parasitol.* **31**, 753–762 (2001).
88. Kumar, S. *et al.* A multilateral effort to develop and test multi-stage, multi-gene DNA vaccines against *Plasmodium falciparum* malaria. *Trends Parasitol.* (in the press).

Acknowledgements

We thank J. Cohen, D. Doolan, J. Epstein, A. Hill, S. Jepsen, N. Tornieporth and S. Hoffman for comments. Funding has been provided by the US Navy Bureau of Medicine, the US Army Military Infectious Disease Research Program, the Office of Naval Research, and USAID. The opinions and assertions herein are those of the authors and are not to be construed as official or as reflecting the views of the US Navy or the US Government.

Plasmodium, human and Anopheles genomics and malaria

Stephen L. Hoffman*, G. Mani Subramanian*, Frank H. Collins† & J. Craig Venter*

*Celera Genomics, 45 West Gude Drive, Rockville, Maryland 20850, USA (e-mail: stephen.hoffman@celera.com)

†University of Notre Dame, Notre Dame, Indiana 46556, USA

The *Plasmodium* spp. parasites that cause malaria are transmitted to humans by *Anopheles* spp. mosquitoes. Scientists have now amassed a great body of knowledge about the parasite, its mosquito vector and human host. Yet this year there will be 300–500 million new malaria infections and 1–3 million deaths caused by the disease. We believe that integrated analyses of genome sequence, DNA polymorphisms, and messenger RNA and protein expression profiles will lead to greater understanding of the molecular basis of vector–human and host–parasite interactions and provide strategies to build upon these insights to develop interventions to mitigate human morbidity and mortality from malaria.

By the end of 2002, the genome sequences of many of the key species in the life cycle of the malaria parasite will be available^{1–3}. This includes: (1) the 25–30 million base pairs (megabase or Mb) of the genomes of *Plasmodium falciparum* — the parasite responsible for more than 95% of all malaria deaths — and *P. yoelii*, a rodent parasite; (2) the 280 Mb of the genome of *Anopheles gambiae*, the most important mosquito vector of *P. falciparum* in sub-Saharan Africa, where malaria causes the most deaths; and (3) the 2,900 and 2,600 Mb of the human and mouse genomes, respectively. Here, we summarize the current status of the *Plasmodium*, human and *Anopheles* genome sequencing projects. We show how a number of fields of biology, based on genomics, enabled and integrated by bioinformatics, and often referred to collectively as ‘functional genomics’, are being brought to bear on malaria research, and how we think these approaches will improve the chances of success in our battle against malaria.

Status of *Plasmodium* spp. sequencing projects

To sequence the genome of *P. falciparum*⁴, a consortium of genome centres and funders was established in 1996, which included The Institute for Genomic Research and Naval Medical Research Center (TIGR/NMRC), Sanger Institute and Stanford University (Table 1). The ~25-Mb *P. falciparum* (clone 3D7) genome, comprising 14 chromosomes and an estimated 5,000 genes, has been sequenced using a whole-chromosome shotgun strategy. Additional sequence information includes the 6-kilobase (kb) mitochondrial genome⁵ and a 35-kb circular DNA that localizes to a novel organelle called the apicoplast⁶. The sequences of chromosomes 2 and 3 were published in 1998 and 1999 respectively^{7,8}, and a draft sequence and annotation of the entire genome will be published in 2002.

Experimental and computational hurdles had to be overcome to complete sequencing and assembly of this genome^{7–9}. Large genomic fragments of *P. falciparum* DNA are not stable in *Escherichia coli*, a problem thought in large part to be due to the skewing of nucleotide content from the 59% adenine and thymine (A+T) of the human genome and 49% of the *E. coli* genome to the 80% found in the *P. falciparum* genome. The presence of A+T

repeats/homopolymers has also made gap closure and accurate assembly difficult. New techniques⁹ and annotation software¹⁰ had to be developed to overcome these problems. Additional resources used to facilitate physical gap closure included use of sequence-tagged site markers derived from end sequences of yeast artificial chromosomes (YACs) previously mapped to the genome, microsatellite markers and a high-resolution linkage map¹¹. Additionally, groups of contigs were ordered on the chromosomes by reference to optical restriction maps¹². The genome-wide, high-resolution linkage map comprises 901 markers that fall into 14 inferred linkage groups with a high rate of uniform meiotic crossover activity¹¹.

Progress in other *Plasmodium* spp. sequencing initiatives is summarized in Table 1. Based on analysis of the *P. falciparum* genome sequencing effort, it was estimated that a three- to fivefold shotgun sequence coverage of the entire genomes of other *Plasmodium* spp. would give 90–95% of the information provided by the significantly more expensive and time-consuming 10–15-fold coverage on a chromosome-by-chromosome basis of the *P. falciparum* genome. A fivefold shotgun sequence of the genome of the rodent parasite *P. yoelii* has been completed, and will be published in 2002. It is anticipated that a threefold shotgun sequence of the genomes of the rodent parasites *P. chabaudi* and *P. berghei* will be completed soon, as will a fivefold coverage of the human parasite *P. vivax*, and of the simian parasite *P. knowlesi*. Additional ongoing efforts are directed at sequencing expressed sequence tags (ESTs; Table 1) of *P. falciparum*, *P. yoelii*¹³, *P. berghei*¹⁴ and *P. vivax*¹⁴. As genomic sequence from other related apicomplexans, such as *Cryptosporidium* and *Toxoplasma*, become available, comparative genomic analyses will enhance understanding of these protozoan pathogens.

To facilitate interpretation and dissemination of genomic sequence data to researchers worldwide, informatics capabilities that integrate sequence data, automated analyses and annotation data emerging from the *P. falciparum* genome project have been developed (ref. 15; and Table 1).

Insights from *P. falciparum* genome data

The biological insights gleaned from publicly available genomic sequence data have already led to the identification of important leads for antimalarial drugs and potential

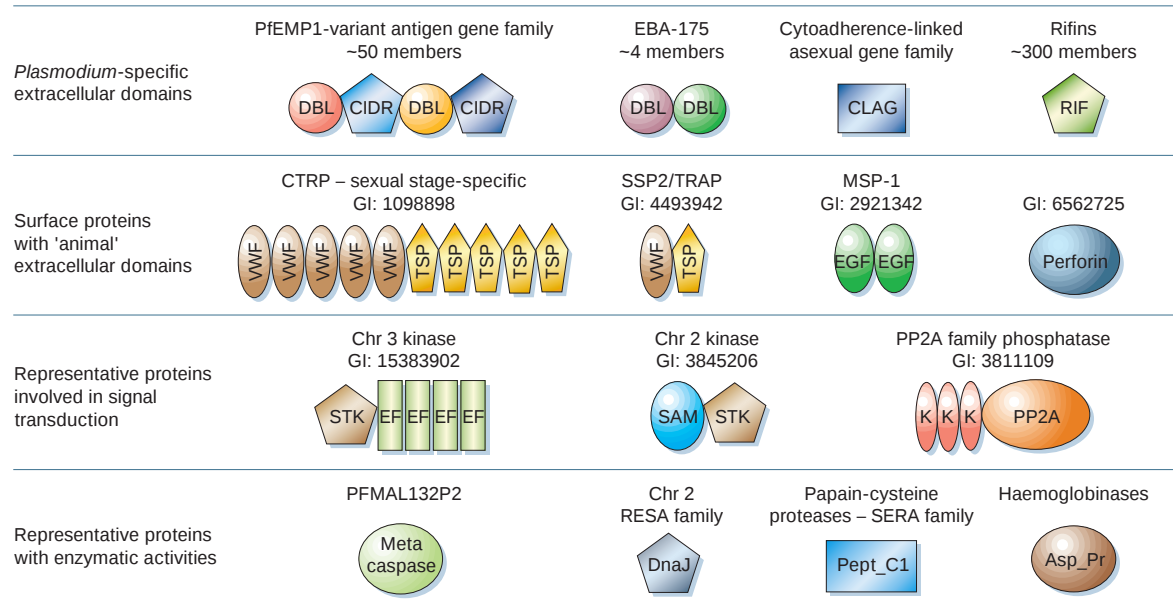


Figure 1 Protein domain architecture in representative *Plasmodium* proteins. Domain name abbreviations: DBL, Duffy-binding-like; CIDR, cysteine-rich interdomain region; CLAG, cytoadherence-linked asexual gene; RIF, Rifins; VWF, von-Willebrand factor A; TSP, thrombospondin; EGF, epidermal growth factor-like; Perforin, membrane-activated complex-perforin; STK, serine/threonine kinase; EF, calcium-binding EF-hand; SAM, sterile alpha-motif; K, kelch repeats; PP2A, PP2A phosphatase; DnaJ, DnaJ family of chaperones; Pept_C1, papain cysteine protease; Asp_Pr, aspartyl protease.

targets for vaccines (refs 16–18; and see Table 2). Furthermore, computational analysis of published genomic sequence data has improved our understanding of the evolutionary and functional adaptations of *P. falciparum*.

Target and lead identification for drugs based on genome sequence data

A rapid demonstration of how genomic sequence facilitates target and lead identification occurred almost as soon as the preliminary, uncompleted sequence of chromosome 2 was posted on the TIGR website. Using homology to known genes in plants and algae, researchers identified several lipid biosynthesis genes in *P. falciparum* and demonstrated their targeting to the apicoplast¹⁹. These genes are attractive drug targets given their critical role in parasite viability and the absence of the type II fatty acid biosynthesis pathway in humans;

at least one major pharmaceutical company has assessed the effect on malaria parasites of compounds known to be active in this pathway. In follow-up, using sequence data on the Sanger website, researchers identified a key enzyme in this pathway — enoyl-acyl-carrier protein (enoyl-ACP) reductase (FabI) — and demonstrated *in vitro* activity against *P. falciparum* using an inhibitor to this enzyme¹⁷.

Another effort was based on unpublished data from chromosome 14 posted on the TIGR website. Researchers used genomic sequence data to identify the enzymes 1-deoxy-D-xylulose-5-phosphate (DOXP) synthase and DOXP reductoisomerase, and provided evidence that isoprenoid biosynthesis in *P. falciparum* depends on the DOXP pathway and is critical for parasite viability¹⁸. They identified two drugs known to be active in this pathway, fosmidomycin and its derivative FR-900089, and demonstrated

Table 1 Current status of *Plasmodium* sequencing projects

<i>Plasmodium</i> species	Sequencing approach	n-Fold coverage	Status	Institution publications*	Anticipated full publication
<i>P. falciparum</i> Chromosomes 2,10,11,14	Whole-chromosome shotgun sequencing	10–15 ×	Completed	TIGR/NMRC ⁷	Mid-2002
<i>P. falciparum</i> Chromosomes 1,3,4,5–9,13	Whole-chromosome shotgun sequencing; includes YAC shotgun sequence	10–15 ×	Near completion	Sanger ⁸	Mid-2002
<i>P. falciparum</i> Chromosome 12	Whole-chromosome shotgun sequencing; includes YAC shotgun sequence	10–15 ×	Completed	Stanford	Mid-2002
<i>P. yoelii</i>	Whole-genome shotgun sequencing	5 ×	Completed	TIGR/NMRC	Mid-2002
<i>P. chabaudi</i>	Whole-chromosome shotgun sequencing	3 ×	In progress	Sanger	?
<i>P. berghei</i>	Whole-chromosome shotgun sequencing	3 ×	In progress	Sanger	Mid-2002
<i>P. knowlesi</i>	Whole-chromosome shotgun sequencing	5 ×	In progress	Sanger	Mid-2002
<i>P. vivax</i>	Whole-genome shotgun sequencing	5 ×	In progress	TIGR/NMRC	End 2002
<i>P. yoelii</i>	Sporozoite cDNA sequencing	>1,900 ESTs	Completed	TIGR/NMRC/NYU ¹³	August 2001
<i>P. vivax</i>	Genome sequence tags (genome survey sequences; GSSs)	>10,000 GSSs	In progress	Univ. Florida ¹⁴	December 2001
<i>P. berghei</i> ANKA clone	EST sequencing; genome sequence tags	>5,000 ESTs >5,000 GSSs	In progress	Univ. Florida ¹⁴	December 2001

*Institutional websites and additional Internet database resources are as follows. *Plasmodium* genome database at TIGR/NMRC: <http://www.tigr.org>; <http://www.nmri.nmrc.navy.mil>. *Plasmodium* genome database at Sanger: http://www.sanger.ac.uk/Projects/P_falciparum/. *Plasmodium* genome database at Stanford: <http://www.sequence.stanford.edu/group/malaria>. Malaria genome tag sequencing projects at University of Florida: <http://parasite.vetmed.ufl.edu/>. The University of Pennsylvania provides access to all of the genome data produced by the *P. falciparum* genome project (PlasmoDB): <http://plasmodb.org/>. WHO/TDR database: <http://www.who.edu.au/MalDB-www/who.html>. National Center for Biotechnology Information Malaria Genetics and Genomics: <http://www.ncbi.nlm.nih.gov/Malaria/>. Metabolic pathways in *Plasmodium*: <http://sites.huji.ac.il/malaria/>.

Table 2 Highlights of data generated from the *Plasmodium* genome project

	Implications on malaria research (diagnostics and drug or vaccine development)
High A + T content ⁷	Technical difficulties in cloning, sequencing and assembly. May confound DNA-based diagnostics.
High percentage of low-complexity proteins ⁷	Variable specificity of antibody reagents. May confound serum-based diagnostics. Need for altered strategies in immunogen production and vaccine design.
Surface protein families (also refer to Fig. 1)	
<i>Plasmodium</i> -specific expansions (vars, rifins, stevors and clags) ^{7,8,26–28}	Central role in cytoadherence and pathology associated with malaria. The antigenic variations in these families pose a huge challenge for vaccine development.
Several protein families with animal-specific extracellular domains ^{7,8,35–39}	Mediate interactions with host receptors and constitute some of the most promising vaccine candidates.
Metabolic enzymes (plastid derived)	
Type II fatty acid biosynthesis ^{17,19}	The cyanobacterial inheritance (absent in humans) makes them attractive as drug targets.
Isoprenoid biosynthesis ¹⁸	Inhibitors of enoyl-ACP reductase have shown <i>in vitro</i> and <i>in vivo</i> activity.
Shikimate pathway ¹⁶	Inhibitors of the DOXP pathway have shown <i>in vitro</i> and <i>in vivo</i> activity. Preliminary evidence supports the importance of this pathway for parasite viability.
Unusual features of regulatory proteins	
Proteases	
Papain family cysteine proteases ⁷	The serine-rich antigen (SERA) family of proteases are related structurally to the cysteine proteases. Candidates for therapeutic (drug or vaccine) development.
Aspartyl proteases ^{7,8,21}	Central role in haemoglobin digestion in the red cell stages. Attractive drug targets.
Metacaspases ²³	Recently described family of caspases (cysteine proteases) that are absent in animals.
DnaJ family chaperones ⁷	Significant expansion of this family suggests diverse roles for this protein domain in the <i>Plasmodium</i> life cycle.
Serine/threonine kinases ^{7,8}	Several Ser/Thr kinases with many containing calcium-binding EF-hands; indicative of a prominent role for calcium mediated signal transduction. Other features of signal transduction are consistent with the early branching of <i>Plasmodium</i> from the eukaryotic radiation.
Likely absence of tyrosine kinases/phosphatases ^{7,8}	
Paucity of transcription factors ^{7,8}	
Genetic map ¹¹	The high-resolution map has greatly helped validate assembly of the genome and markers to identify genetic correlates of drug resistance.
Genetics of drug resistance ^{77,78}	Genetic and biochemical approaches have been used to identify mutations in plasmodial genes to dissect the molecular basis for drug resistance. The comprehensive catalogue of these mutations will enhance efforts in monitoring the spread of drug resistance.
Chloroquine	
Dihydrofolate reductase antagonists	
Stage-specific RNA expression	
Asexual versus sexual stages ⁴³	Identified several stage-specific proteins, several of which are vaccine candidates.
Sporozoite stages (<i>P. yoelii</i>) ¹³	Use of syntenic relationships to identify the orthologous gene in <i>P. falciparum</i> . Provides a set of genes to further investigate as vaccine candidates.

activity *in vitro* against *P. falciparum*, including against multidrug-resistant *P. falciparum*, and *in vivo* against *P. vinckei* in mice¹⁸.

The value of genomic sequence data and comparative analyses to identify new drug targets was shown recently with the identification in *P. falciparum* of the metal-dependent RNA triphosphatase protein family, members of which are crucial in mRNA cap formation and eukaryotic gene expression²⁰. The structure of the active site and catalytic mechanism of this protein family in *P. falciparum* and fungi are completely different from the RNA triphosphatase domain of the metazoan (human) capping enzymes, and metazoans encode no identifiable homologues of the fungal or *Plasmodium* RNA triphosphatases. The structural similarity between the plasmodial and the fungal RNA triphosphatases raises the exciting possibility of achieving antifungal and antimalarial activity with a single class of mechanism-based inhibitors²⁰. Protein families that are expanded in *P. falciparum* and are considered attractive drug targets include the cysteine and aspartyl class of proteases, with both classes constituting a parasite adaptation to its need for haemoglobin digestion^{7,21,22}. Of special interest is the representation of several members of the recently described metacaspase family of cysteine proteases in *Plasmodium*²³. The absence of this class of proteases in humans makes the *P. falciparum* metacaspases particularly attractive as drug targets.

Target identification for vaccines based on genome sequence data

P. falciparum proteins that traffic to the erythrocyte surface are important in the pathogenesis of malaria, and are potentially important targets for vaccine development²⁴. Genomic sequence data have demonstrated the extent of paralogous expansion of genes encoding such parasite proteins. Several members of these multi-gene families are being considered as vaccine candidates given their prominent role in malarial pathogenesis (refs 24,25; and Table 2, Fig. 1). Prominent among these are the variant gene families clustered in the subtelomeric regions of the *P. falciparum* genome (*var* genes encoding *P. falciparum* erythrocyte membrane protein (PfEMP), and *rifin*, *stevor* and *clag* genes)^{7,26–28} and *P. vivax* genome (*vir* genes)²⁹. Furthermore, the availability of genomic sequence data has provided invaluable clues into the regulation of the large variant-antigen PfEMP gene family^{7,8,30}. Other attractive targets for vaccines are parasite proteins important for parasite invasion of erythrocytes

(see reviews in this issue by Richie and Saul, pages 694–701, and Miller *et al.*, pages 673–679). Here again, genomic sequence data have enabled the identification and characterization of paralogues of the Duffy-binding-like domain-containing, EBA-175-like gene family members^{31,32}, and the rhoptry³³ and reticulocyte binding protein³⁴ families involved in red cell invasion.

Although most of these merozoite and erythrocyte surface proteins have plasmodial-specific domains, it is of interest to note that there are several examples of plasmodial proteins with extracellular binding domains that are found predominantly in animal genomes (Table 2, Fig. 1). These include the epidermal growth factor domain-containing merozoite surface proteins and Pfs25/28 (*P. falciparum* gametocyte surface protein) that have homologues in other plasmodium species^{35,36}. Likewise, thrombospondin and von-Willebrand factor A domain-containing proteins such as circumsporozoite protein (CSP), sporozoite surface protein 2/thrombospondin-related adhesive protein (SSP2/TRAP) and CSP-TRAP-related protein (CTRP) are involved in the gametocyte and sporozoite stages of the life cycles^{37,38}. Many of these proteins are central to initiating infection of the host, elicit a protective immune response, and are being developed as vaccine candidates (see reviews by Richie and Saul and by Miller *et al.* in this issue). The complete genomic sequence allows for rapid identification of all members of these families and their investigation as potential targets for vaccine development. Other interesting examples of animal extracellular adhesion domains detected in predicted plasmodium proteins include the LCCL (domain identified in *Limulus* factor C, Coch-5b2 and LGL1) module³⁹ and perforin domain-containing proteins⁴⁰ in the *P. falciparum* genome.

Finally, over 40% of predicted proteins in the genome are predominantly low-complexity or non-globular proteins with several instances of low-complexity inserts in the midst of globular protein domains⁷. This may reflect an adaptive mechanism of the parasite to evade an antibody-driven immunity and may impact vaccine design to improve specificity of antibody response.

mRNA and protein expression to identify drug and vaccine targets

Genomic sequence data and new technologies have provided the foundation for studies to determine mRNA and protein expression at different stages of the parasite life cycle⁴¹. Over 3,000 random inserts

from a *P. falciparum* mung bean nuclease genomic library were used in a DNA microarray to identify differences in gene expression between the asexual blood-stage trophozoite and the sexual blood-stage gametocyte form of the parasite⁴². These experiments increased the list of stage-specific transcripts in *P. falciparum* by an order of magnitude, provided potential targets for developing drugs and vaccines, and yielded clues as to how to explore differences in the metabolic machinery of the parasite as it transits from humans to mosquitoes⁴³. A DNA microarray based on genes from chromosomes 2, 3, 12 and 14 was used to assess differential parasite gene responses to antimalarial drugs in drug-sensitive and drug-resistant strains at different times during the asexual erythrocytic stage of the life cycle (ref. 41; and A. Whitney, unpublished data). A different approach, based on sequencing of a *P. yoelii* sporozoite complementary DNA library, helped identify genes expressed in the sporozoite stage of the life cycle¹³. In addition to identifying over 1,300 new sporozoite-expressed genes, this study documented the sporozoite-stage expression of potentially important surface molecules.

Work has also begun to identify proteins expressed at specific stages of the parasite life cycle⁴¹; data that will be especially important for vaccine development. There are multiple methods available for conducting such studies, but the most progress to date has used recent technological advances in liquid chromatography–mass spectrometry (LC-MS), wherein mass spectral fingerprints that are experimentally derived may be matched to computationally generated profiles derived from the genomic sequence (refs 41,44; and D. Carucci, personal communication). The availability of an accurately assembled and annotated complete genome is critical for the success of these ongoing efforts.

Status of the human genome project

The 2,900-Mb human genome, with approximately 30,000 predicted protein-coding genes^{1,2} and a genome-wide catalogue of several million single nucleotide polymorphisms (SNPs), is now available for researchers to better define the genetic determinants of human disease^{1,45}. Furthermore, ongoing efforts at haplotype analysis will help discern disease-related SNPs from the background of irrelevant genetic variation⁴⁶. Likewise, efforts are underway to obtain portraits of gene and protein expression profiles from human tissue from a wide range of anatomic compartments in the body during different states of health and disease and over time. These projects are in large part dependent on the availability of the complete genomic sequence. The relevance of the human genome sequence to human disease and medicine has been discussed in a recent publication⁴⁷.

Relevant insights from the human genome project

Human genetic variations are one of the principal determinants of susceptibility to many common infectious diseases⁴⁸. Malaria was one of the first infectious diseases to be studied extensively and many susceptibility and resistance loci have been identified over the past few decades⁴⁹. The most prominent human polymorphism that protects against severe malaria is the sickle cell trait, although other haemoglobinopathies or blood group antigen variants have been shown to influence clinical outcomes by altering the efficiency of invasion of erythrocytes by the parasites or the development of the parasites within erythrocytes.

Other determinants of clinical severity include receptor polymorphisms in genes encoding proteins such as intercellular adhesion molecule 1 (CD54) and CD36, which are involved in the sequestration of infected red cells in vascular endothelium, or polymorphisms in the human tumour-necrosis factor- α gene promoter that are associated with cerebral malaria^{50–52}. Although a dozen or more specific susceptibility determinants have been defined so far, including both structural and regulatory polymorphisms of red cell proteins and genes of the immune system, it is most likely that many more are yet to be discovered⁴⁹. It must also be emphasized that efforts made so far do not represent a methodical and comprehensive

survey of the genome. As the human genome sequence and the extensive catalogue of SNPs have been publicly available only since early 2001, we have yet to reap the full benefits of this tremendous resource for malaria research.

Status of the *Anopheles* genome project

The *A. gambiae* genome project is funded largely by the National Institute of Allergy and Infectious Diseases and the French Government, with the World Health Organization/Tropical Disease Research programme providing a coordinating role and assisting in database development. Celera Genomics and Génoscope have recently completed a shotgun sequencing effort (10 $\times$ coverage), which has been assembled by Celera. Two independent annotations, one by Celera and the other by the European Bioinformatics Institute using Ensembl should be completed by early 2002. The annotated genome may be presented publicly as early as March 2002 either through Ensembl or through an adapted version of the *Drosophila* GdFly. End sequences of two different *A. gambiae* bacterial artificial chromosome (BAC) libraries (>10 $\times$ coverage), the physical map location of ~2,000 BAC clones, perhaps as many as 80,000 ESTs, and several sequenced and annotated genomic DNA contigs (100–500+ kb) will facilitate assembly and annotation. Celera Genomics, the European Molecular Biology Laboratory, Génoscope, the Institute of Molecular Biology and Biotechnology in Crete, the Institut Pasteur in Paris, TIGR, and the Universities of Iowa, Rome and Notre Dame are contributing.

All genomic libraries used in this project have been made from a strain of *A. gambiae* (PEST) that has a sex-linked, pink-eye mutation⁵³ and a standard chromosome arrangement⁵⁴. The shotgun sequence has been obtained from plasmid libraries with 2-kb, 10-kb and 50-kb inserts produced at Celera. All BAC libraries, sequenced cDNAs and the end-sequenced 10-kb plasmid clones will be archived at the American Type Culture Collection as part of the National Institutes of Health-supported Malaria Research and Reference Reagents Resource Center (<http://www.malaria.mr4.org/mr4pages/index.html>).

Relevant insights from the *Anopheles* genome project

The 120-Mb genome of fruitfly, *Drosophila melanogaster*, provided the first complete invertebrate insect genome sequenced using the shotgun sequence and assembly approach⁵⁵. The completion of the 280-Mb *Anopheles* genome will undoubtedly enhance our understanding of the evolutionary adaptations of mosquitoes since their divergence from the fly approximately 200 million years ago, including traits such as the physiology of blood feeding and blood meal digestion and behaviours associated with the selection of humans as the blood meal source and oviposition in aquatic sites produced by human agricultural activity. Most of the specific EST efforts in this project are focused on mosquito tissues with which malaria parasites interact⁵⁶, including the midgut, haemocytes, fat body and salivary glands, or on genes expressed in tissues such as the head and antennae that may mediate behaviours such as host selection and mating. A number of current research efforts are motivated by an interest in the mosquito's response to infection with the malaria parasite, as well as the basis whereby several selected strains of *A. gambiae* are able to mount a defensive response to *Plasmodium*^{57,58}.

Comparative analysis of the *Drosophila*, *Anopheles* and human genomes has already provided several practical insights. The shared aspects of innate immunity between the two insects and humans have been confirmed by computational^{1,59} and experimental evidence^{56,60}. Of special interest are the specific adaptations of haematophagous insects (including *Anopheles*) that promote transmission by enhancing blood feeding. These include the convergent evolution of proteins that interact extensively with mammalian proteins which regulate haemostasis and the complement cascade⁶¹. On a larger comparative scale, preliminary examinations of conserved gene arrangement between *Drosophila* and *A. gambiae* suggest that substantial gene order arrangement may be preserved at a scale of up to

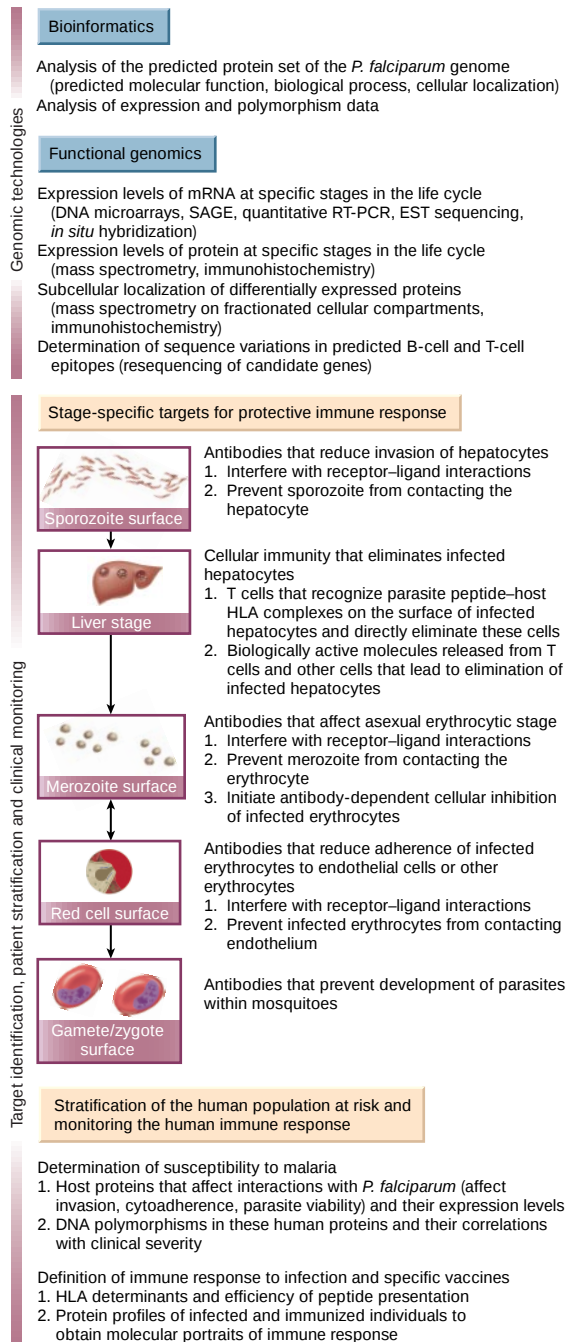


Figure 2 Malaria 'vaccinomics'. The potential impact of genomics and functional genomics on identification of new functionally or anatomically important targets for malaria vaccine development, and optimizing selection of human populations for immunization with specific vaccines.

several hundred kilobases, and very large blocks of most chromosome arms may be shared⁶². The ability to compare mosquito and *Drosophila* genes in terms of sequence similarity and location will be important in guiding researchers towards an understanding of mosquito gene function.

What does genomics offer for the control of malaria?

There is currently no vaccine for malaria, no optimal way to sustainably reduce or eliminate contact between humans and infective mosquitoes, a poor system for early diagnosis and recognition of those at greatest risk for developing severe disease, and a disturbing,

ever increasing resistance to the drugs used to treat malaria. We believe that genomics and related disciplines provide the scientific foundation for improving our chances to address and solve each of these problems⁶³.

Vaccine development

Effective vaccines elicit immune responses that destroy the infectious agent or the host cells in which they reside, or inhibit a function critical for its survival. Immune responses against *P. falciparum* can protect through both mechanisms. Immunization with radiation-attenuated sporozoites protects more than 90% of recipients against experimental challenge for at least 10 months⁶⁴. Protection is thought to be mediated primarily by T cells against peptides from parasite proteins expressed in infected hepatocytes (function independent), although antibodies that reduce sporozoite invasion of hepatocytes (function inhibiting) also have a role (refs 65–67; and Fig. 2). In areas of Africa with the most intense malaria transmission, children who survive to the age of 7–10 rarely develop life-threatening *P. falciparum* infections. They become infected frequently, but their immune systems limit the infections, thereby preventing severe disease. Antibodies against parasite proteins expressed on erythrocytes, which prevent sequestration in the microcirculation (function inhibiting)²⁵, and on the surface of merozoites, which prevent invasion of erythrocytes (function inhibiting)⁶⁸ or initiate antibody-dependent cellular inhibition⁶⁹ activity (function independent), are thought to be primary in this 'naturally acquired' protective immunity.

During the past 20 years there has been considerable work to develop subunit vaccines that provide protection after exposure to irradiated sporozoites and naturally acquired immunity (see review in this issue by Richie and Saul, pages 694–701). Such a vaccine has yet to be developed. There are numerous potential explanations for the lack of current success. One is that exposure to the whole parasite elicits a more potent, protective immune response than do the subunit vaccines tested thus far. However, in the case of people protected by the irradiated sporozoite vaccine, or by naturally acquired immunity, T-cell and antibody responses are generally modest, and in many instances lower than after subunit vaccination. It is more likely that immunization with only a few parasite proteins cannot duplicate the immunity elicited by exposure to a parasite that has thousands of proteins; modest immune response against tens, hundreds or thousands of parasite proteins may be additive or synergistic. In the case of naturally acquired immunity, this 'breadth' would be expanded by exposure to so many polymorphic strains of *P. falciparum*. If this is the case, then the malaria genome and SNP projects may provide the essential foundation for duplicating this whole-organism immunity.

Analysis of the sequence of the *P. falciparum* genome has dramatically expanded our knowledge regarding the paralogous expansion in the genome of parasite proteins expressed on the erythrocyte surface (Table 2, Fig. 1). But genomics on its own will not be enough⁷⁰. Immune responses are directed against proteins, not genes, and the parasite expresses different proteins at different stages of its life cycle⁷⁰.

We believe that the first step in duplicating whole parasite-induced protective immunity is to catalogue expression of proteins at each stage of the parasite life cycle. The first step to accomplishing this could involve generating antibodies to every parasite protein and determining their subcellular localization⁷⁰. This cannot be done rapidly and without great expense, so other approaches have been initiated. Assessment of mRNA and/or protein expression at different stages of the parasite life cycle would provide comprehensive data. Unfortunately, neither is 'the best' way. Immune responses recognize proteins, not mRNA, so gene expression profiles must be confirmed at the protein level. Furthermore, there are many instances (for example, alternative splice sites or post-translational modifications) in which mRNA expression data will not be adequate. The inability to culture infective *P. falciparum* sporozoites and the poor performance of hepatocyte cultures make it difficult (sporozoites) or impossible (liver-stage parasites) to acquire enough parasite material for

conventional, non-amplified gene chip or DNA microarray analysis of these critical stages of the life cycle. For these reasons there has been increasing emphasis on LC-MS methods for establishing differential protein expression profiles at each stage of the life cycle⁴¹.

We anticipate the generation of comprehensive databases that catalogue stage-specific expression of all genes and proteins in the *P. falciparum* genome, and the polymorphisms in the DNA sequences of these genes in field isolates. Concurrent efforts to establish the functional importance of each gene, and the significance of their protein–protein interactions will take longer. Nonetheless, characterization of genes, their transcripts, and the proteins they encode will provide the foundation for the development of effective vaccines.

In subsequent sections we address how genomics-based data may be helpful in focusing vaccination strategies. This will include the need to identify patients at high risk, as well as a ‘systems biology/immunology’ approach to monitor immune response, wherein molecular kinetic ‘portraits’ of individuals should provide a more comprehensive indication of the response to infection and immunization than do currently available cross-sectional single immune response measurements (Fig. 2, Table 3). Significant difficulties lie ahead in developing vaccines that duplicate the protection provided by immunization with irradiated sporozoites or naturally acquired immunity (or that work through other mechanisms), delivering those vaccines optimally to the individuals who need them most, and effectively monitoring responses to those vaccines. In addition to increased understanding of the target of immune responses (the parasite proteins and epitopes within those proteins), work in vaccinology that takes advantage of the human genome project to develop improved methods for maximizing magnitude, quality and longevity of protective immune responses will be fundamental to fielding effective malaria vaccines.

Reducing contact between humans and infective mosquitoes

The sequence of the *A. gambiae* genome will accelerate development of new malaria control tools to supplement existing insecticides and bednet strategies, not only by advancing current research efforts, but also, and perhaps most important, by opening lines of research not currently envisaged. The dire need for new vector-related tools and strategies for malaria control is obvious given that most of the world’s successful malaria control programmes have utilized primarily transmission-reduction tools such as insecticides and bednets⁷¹.

Preserving the efficacy of insecticides in the face of resistance

The most important vector-targeted control strategies today involve insecticide-impregnated bednets. Unfortunately, resistance to the

pyrethroid insecticides used in bednets has now been recorded in *A. gambiae* populations in both east and west Africa as well as in *A. funestus* (another major vector in Africa) from southern Africa⁷¹. One of the main causes of resistance is upregulation of groups of insecticide-detoxifying enzymes such as oxidases, esterases and glutathione *S*-transferases⁷². The *A. gambiae* genome will enable rapid identification of the genes involved and the molecular changes that lead to resistance, information that may guide in the selection and/or development of alternative insecticides. It will also provide tools that can be used in population studies to monitor the emergence and spread of important traits like insecticide resistance.

Increasing the efficiency of vector-targeted control efforts

A. gambiae is a member of a complex of what are now recognized as seven related but genetically and ecologically distinct species⁷¹. Decades of field work in Africa, where the vector was identified as *A. gambiae sensu lato*, have produced a picture of *A. gambiae* ecology that is probably the composite of data from at least two or even more members of this complex. Recent studies in west Africa suggest *A. gambiae sensu stricto* may comprise at least two and possibly as many as five distinct, reproductively isolated cytogenetic forms⁷³. An immediate outcome of the *A. gambiae* genome project should be development of simple, DNA-based diagnostics for these different forms that can be used to more fully understand cytotype-specific differences in vector behaviour, ecology and pathogen transmission.

Development of a genetic control strategy for *A. gambiae*

The *A. gambiae* genome project builds on a decade of progress in developing tools for the genetic control of malaria transmission⁷⁴. Robust germline-transformation tools now exist for *Anopheles* mosquitoes⁷⁵. Significant progress has been made in the study of parasite development in the *Anopheles* vector^{58,71}. For example, it has now been shown that normal parasite development can be almost totally blocked in *Anopheles* mosquitoes transformed with a construct that binds to both midgut and salivary glands (M. Jacobs-Lorena, personal communication). Considerable progress has also been made in studies of *A. gambiae* population structure, using tools that emerged from laboratory cloning and mapping efforts⁷¹. The *A. gambiae* genome project will help solve the problems that must be overcome before even a limited genetic-control field trial can be implemented.

Early diagnosis and predicting risk of developing severe malaria

To achieve the goal of reducing morbidity and mortality caused by malaria, it will be imperative to target expensive resources to those at

Table 3 Integrative analysis of malaria using genome-based computational and experimental approaches

Genomic technologies	Vaccine development	Drug development	Early diagnosis and predicting risk of developing severe malaria	Reducing and eliminating contact between humans and infective mosquitoes
Genomic sequence analysis Predict gene structure and function Comparative genomics	Target identification: predict surface proteins; identify homologues across <i>Plasmodium</i> spp.; mouse models of malaria to identify and validate vaccine candidates.	Target identification: identify new or divergent members in a biological pathway in <i>Plasmodium</i> ; structure-based rational drug design; mouse models of malaria to identify drug targets and to model efficacy of new therapies.	Comparative analysis across <i>Plasmodium</i> spp. to define determinants of drug resistance; mouse models of malaria to identify genetic determinants of resistance and clinical severity.	Target identification: identify new or divergent members in a biological pathway in <i>Anopheles</i> (insecticide)
Resequencing Catalogue genetic variations in <i>Plasmodium</i> , <i>Anopheles</i> and human populations	Evaluate polymorphisms in vaccine epitopes to optimize vaccine design.	Evaluate polymorphisms in genes that are drug targets (in <i>Plasmodium</i>) or in drug modifiers (in humans)	Identify polymorphisms in <i>Plasmodium</i> genes that impact drug resistance, and in human genes that impact susceptibility to severe malaria.	Evaluate polymorphisms in genes in <i>Anopheles</i> that will allow detection of resistance to insecticides and efficiency of transmission.
Differential expression RNA (microarray) Protein (LC-MS) Metabolite (LC-MS)	Target identification of proteins that are relevant for T-cell or B-cell vaccines based on stage-specific differential expression.	Target identification and validation based on stage-specific expression	In <i>Plasmodium</i> , identify surrogate markers for diagnosis and predictors of clinical response to therapy; in humans, monitor immune response profiles and severity of illness.	Identify genes in <i>Anopheles</i> that modify the innate immune response to <i>Plasmodium</i> and resistance to insecticides.
Protein–interaction maps Yeast two-hybrid genetic screens Mass spectrometry Antibody or protein chips	Identification of receptor–ligand interactions to define critical binding motifs for vaccine design.	Increased numbers of potential drug targets based on the understanding of biological pathways.	Not applicable.	Increased numbers of potential insecticide targets and targets to inhibit sporogonic development based on the understanding of biological pathways.

highest risk of suffering severe disease and death. We believe that genomics-based efforts may facilitate this approach.

Early diagnosis of infection

Conventional approaches for diagnosing *P. falciparum* infection include microscopic examination of blood smears, the use of dipsticks that assay *P. falciparum* histidine-rich protein 2, and the less practical amplification of *P. falciparum* genes using the polymerase chain reaction (PCR)⁷⁶. Gene expression and proteomics profiling studies will probably provide additional markers. We speculate that several of these protein- or metabolite-based markers could be adapted in antibody-based assays to enhance detection as well as provide surrogate markers of drug resistance and clinical outcomes (Table 3).

Rapid identification and monitoring of drug resistance

It is imperative that drug-resistant strains of *P. falciparum* be detected when they first emerge. Genetic and genomic sequence has been the basis of PCR-based screens against individual *P. falciparum* genes used to document and monitor resistance to chloroquine and sulphamethoxazole treatment regimens^{77,78}. The genome-wide, high-resolution linkage map¹¹ is complementary to sequence data and will lead to determination of most drug-resistant genes and their polymorphisms. The capacity to detect *P. falciparum* metabolite and protein levels during the erythrocytic stages of its life cycle using mass spectrometry should lead to the development of clinically relevant, rapid methods for determining drug resistance, and early recognition of failing therapy.

Identification of individuals at risk of developing severe disease

Case-control and association studies that have been used successfully to dissect the genetics of human disease will benefit significantly from the genomic sequence and availability of genetic markers⁴⁸. Additionally, the recent completion of the mouse genome and the comparative genomic map of the two species will complement ongoing efforts to dissect determinants of host susceptibility and immune response in murine models of malaria^{3,79}. Future efforts aimed at identifying genetic correlates in populations at risk with biological correlates of parasite invasion, parasite development, cytoadherence efficiency and clinical outcomes will be critical in identifying at-risk populations to better direct vaccination and chemoprophylaxis strategies (Table 3, Fig. 2). Similar efforts are needed to develop pharmacogenetics in malaria to improve predictors of drug efficacy and toxicity⁸⁰. We foresee a time when such data will be used to direct allocation of resources as well as establish appropriate diagnostic, therapeutic and prevention guidelines in malaria control programmes.

Drug development and new therapeutics

Advances in proteomics will be crucial in identifying differential expressed proteins, which along with comparative genomic analysis, utilization of protein interaction maps and an understanding of metabolic pathways, will help identify and prioritize targets for therapeutic intervention (Table 3, Fig. 2). The ability to target regulatory regions as a therapeutic strategy may now be feasible given developments in computational and experimental approaches to identify potential regulatory regions in *P. falciparum* genes³⁰, and highly selective methodologies that are able to target these highly AT-rich segments⁸¹.

Recent advances in gene-targeting technologies to manipulate *P. falciparum* genes as well as those in other rodent *Plasmodium* species will significantly aid target validation^{82,83}. Of immediate interest will be *Plasmodium* proteins that are amenable to interrogation using currently available small-molecule design platforms and target-family chemical libraries, in addition to being sufficiently divergent or distinct from their human homologue. Structure-based modelling, the identification of critical catalytic and substrate-recognition motifs, polymorphism detection to target invariable regions, and the ability to predict determinants of drug resistance will provide the robust framework to rationalize and streamline the development

of the next generation of therapeutics. Knowledge can be gleaned from the recent success in developing small-molecule inhibitors of the *Trypanosoma cruzi* cysteine proteases⁸⁴, and the *P. falciparum* homologues might also be attractive targets²². Equally tantalizing are the metacaspase family of cysteine proteases²³ that may provide a new class of drug targets (absent in humans) that are amenable to high-throughput screening and the rational drug design process.

An integrated, systems-biology approach in malaria

There are numerous views on how to solve the problems of malaria^{63,85}. We have outlined our perspective on how genomics and genomics-dependent science will be used to develop antimalarial vaccines; reduce or eliminate contact between humans and infective mosquitoes; rapidly diagnose and recognize those at greatest risk for developing severe disease; and identify and treat multidrug-resistant malaria. However, we have not discussed the integration of these efforts, or their integration with complementary fields of biomedicine.

The availability of genomic sequence data, use of sequence-based, high-throughput technologies, and advances in bioinformatics to analyse and interpret genomic data will ultimately provide an integrated picture of malarial biology, pathogenesis and epidemiology. We anticipate a time when we will routinely measure real-time biological profiles, including molecular phenotypic expression and genetic polymorphisms of the vector, host and parasite, and clinical outcomes (Table 3, Fig. 2), to create molecular kinetic portraits. Such complex biological systems have properties that cannot be predicted *a priori*. The development of mathematical modelling methods to describe biological systems at the molecular and/or population level, and predict behaviours of these systems in response to various virtual treatments, will be crucial in developing a new generation of interventions to attack malaria. The promise of a genome-sequence-based platform thereby lies in providing an integrated reconstruction of the spectrum of molecular and cellular interactions among parasite, vector and human host, with the ultimate goal of eradicating malaria. □

- Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
- Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Marshall, E. Genome sequencing. Celera assembles mouse genome: public labs plan new strategy. *Science* **292**, 822 (2001).
- Hoffman, S. L. *et al.* Funding for malaria genome sequencing. *Nature* **387**, 647 (1997).
- Feagin, J. E., Gardner, M. J., Williamson, D. H. & Wilson, R. J. The putative mitochondrial genome of *Plasmodium falciparum*. *J. Protozool.* **38**, 243–245 (1991).
- Wilson, R. J. *et al.* Complete gene map of the plasmid-like DNA of the malaria parasite *Plasmodium falciparum*. *J. Mol. Biol.* **261**, 155–172 (1996).
- Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126–1132 (1998).
- Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400**, 532–538 (1999).
- Gardner, M. J. The genome of the malaria parasite. *Curr. Opin. Genet. Dev.* **9**, 704–708 (1999).
- Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24–31 (1999).
- Su, X. *et al.* A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286**, 1351–1353 (1999).
- Lai, Z. *et al.* A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genet.* **23**, 309–313 (1999).
- Kappe, S. H. *et al.* Exploring the transcriptome of the malaria sporozoite stage. *Proc. Natl Acad. Sci. USA* **98**, 9895–9900 (2001).
- Carlton, J. M. *et al.* Profiling the malaria genome: a gene survey of three species of malaria parasite with comparison to other apicomplexan species. *Mol. Biochem. Parasitol.* **118**, 201–210 (2001).
- The Plasmodium Genome Database Collaborative. PlasmoDB: an integrative database of the *Plasmodium falciparum* genome. Tools for accessing and analyzing finished and unfinished sequence data. *Nucleic Acids Res.* **29**, 66–69 (2001).
- Roberts, F. *et al.* Evidence for the shikimate pathway in apicomplexan parasites. *Nature* **393**, 801–805 (1998).
- Surolia, N. & Surolia, A. Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of *Plasmodium falciparum*. *Nature Med.* **7**, 167–173 (2001).
- Jomaa, H. *et al.* Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science* **285**, 1573–1576 (1999).
- Waller, R. F. *et al.* Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **95**, 12352–12357 (1998).
- Ho, C. K. & Shuman, S. A yeast-like mRNA capping apparatus in *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **98**, 3050–3055 (2001).
- Francis, S. E., Sullivan, D. J. Jr & Goldberg, D. E. Hemoglobin metabolism in the malaria parasite *Plasmodium falciparum*. *Annu. Rev. Microbiol.* **51**, 97–123 (1997).

22. Singh, A. & Rosenthal, P. J. Comparison of efficacies of cysteine protease inhibitors against five strains of *Plasmodium falciparum*. *Antimicrob. Agents Chemother.* **45**, 949–951 (2001).
23. Uren, A. G. *et al.* Identification of paracaspases and metacaspases: two ancient families of caspase-like proteins, one of which plays a key role in MALT lymphoma. *Mol. Cell* **6**, 961–967 (2000).
24. Brown, G. V. & Rogerson, S. J. in *Malaria Vaccine Development: a Multi-Immune Response Approach* (ed. Hoffman, S. L.) 145–166 (ASM Press, Washington DC, 1996).
25. Duffy, P. E., Craig, A. G. & Baruch, D. I. Variant proteins on the surface of malaria-infected erythrocytes—developing vaccines. *Trends Parasitol.* **17**, 354–356 (2001).
26. Su, X. Z. *et al.* The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell* **82**, 89–100 (1995).
27. Kyes, S. A., Rowe, J. A., Kriek, N. & Newbold, C. I. Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **96**, 9333–9338 (1999).
28. Trenholme, K. R. *et al.* *clag9*: A cytoadherence gene in *Plasmodium falciparum* essential for binding of parasitized erythrocytes to CD36. *Proc. Natl Acad. Sci. USA* **97**, 4029–4033 (2000).
29. del Portillo, H. A. *et al.* A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature* **410**, 839–842 (2001).
30. Deitsch, K. W., Calderwood, M. S. & Welles, T. E. Cooperative silencing elements in var genes. *Nature* **412**, 875–876 (2001).
31. Mayer, D. C., Kaneko, O., Hudson-Taylor, D. E., Reid, M. E. & Miller, L. H. Characterization of a *Plasmodium falciparum* erythrocyte-binding protein paralogous to EBA-175. *Proc. Natl Acad. Sci. USA* **98**, 5222–5227 (2001).
32. Narum, D. L., Fuhrmann, S. R., Luu, T. & Sim, B. K. A novel *Plasmodium falciparum* erythrocyte binding protein-2 (EBP2/BAEBL) involved in erythrocyte receptor binding. *Mol. Biochem. Parasitol.* (in the press).
33. Preiser, P. R., Jarra, W., Capod, T. & Snounou, G. A rhoptry-protein-associated mechanism of clonal phenotypic variation in rodent malaria. *Nature* **398**, 618–622 (1999).
34. Trigila, T. *et al.* Identification of proteins from *Plasmodium falciparum* that are homologous to reticulocyte binding proteins in *Plasmodium vivax*. *Infect. Immun.* **69**, 1084–1092 (2001).
35. Black, C. G., Wu, T., Wang, L., Hibbs, A. R. & Coppel, R. L. Merozoite surface protein 8 of *Plasmodium falciparum* contains two epidermal growth factor-like domains. *Mol. Biochem. Parasitol.* **114**, 217–226 (2001).
36. Kaslow, D. C. *et al.* A vaccine candidate from the sexual stage of human malaria that contains EGF-like domains. *Nature* **333**, 74–76 (1988).
37. Sultan, A. A. *et al.* TRAP is necessary for gliding motility and infectivity of plasmodium sporozoites. *Cell* **90**, 511–522 (1997).
38. Templeton, T. J., Kaslow, D. C. & Fidock, D. A. Developmental arrest of the human malaria parasite *Plasmodium vivax* within the mosquito midgut via CTRP gene disruption. *Mol. Microbiol.* **36**, 1–9 (2000).
39. Trexler, M., Banyai, L. & Patthy, L. The LCCL module. *Eur. J. Biochem.* **267**, 5751–5757 (2000).
40. Ponting, C. P. Chlamydial homologues of the MACPF (MAC/perforin) domain. *Curr. Biol.* **9**, R911–R913 (1999).
41. Carucci, D. J. & Hoffman, S. L. The Malaria Genome Project: from sequence to functional analysis. *Nature Med.* **6** <http://www.nature.com/nm/special_focus/malaria/commentaries/carucci.html> (2000).
42. Hayward, R. E. *et al.* Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. *Mol. Microbiol.* **35**, 6–14 (2000).
43. Hayward, R. E. *Plasmodium falciparum* phosphoenolpyruvate carboxykinase is developmentally regulated in gametocytes. *Mol. Biochem. Parasitol.* **107**, 227–240 (2000).
44. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**, 837–846 (2000).
45. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
46. Stephens, J. C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489–493 (2001).
47. Subramanian, G., Adams, M. D., Venter, J. C. & Broder, S. Implications of the human genome for understanding human biology and medicine. *J. Am. Med. Assoc.* **286**, 2296–2307 (2001).
48. Hill, A. V. The genomics and genetics of human infectious disease susceptibility. *Annu. Rev. Genomics Hum. Genet.* **2**, 373–400 (2001).
49. Kwiatkowski, D. The molecular genetic approach to malarial pathogenesis and immunity. *Parasitologia* **41**, 233–240 (1999).
50. Knight, J. C. *et al.* A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria. *Nature Genet.* **22**, 145–150 (1999).
51. Fernandez-Reyes, D. *et al.* A high frequency African coding polymorphism in the N-terminal domain of ICAM-1 predisposing to cerebral malaria in Kenya. *Hum. Mol. Genet.* **6**, 1357–1360 (1997).
52. Pain, A. *et al.* A non-sense mutation in Cd36 gene is associated with protection from severe malaria. *Lancet* **357**, 1502–1503 (2001).
53. Benedict, M. Q. & Chang, H. Rapid isolation of anopheline mosquito eye-colour mutants, based on larval colour change. *Med. Vet. Entomol.* **10**, 93–96 (1996).
54. Mukabayire, O. & Besansky, N. J. Distribution of T1, Q, Pegasus and mariner transposable elements on the polytene chromosomes of PEST, a standard strain of *Anopheles gambiae*. *Chromosoma* **104**, 585–595 (1996).
55. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
56. Dimopoulos, G. *et al.* *Anopheles gambiae* pilot gene discovery project: identification of mosquito innate immunity genes from expressed sequence tags generated from immune-competent cell lines. *Proc. Natl Acad. Sci. USA* **97**, 6619–6624 (2000).
57. Collins, F. H. *et al.* Genetics in the study of mosquito susceptibility to *Plasmodium*. *Parasitologia* **41**, 163–168 (1999).
58. Zheng, L. *et al.* Quantitative trait loci for refractoriness of *Anopheles gambiae* to *Plasmodium cynomolgi* B. *Science* **276**, 425–428 (1997).
59. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
60. Schneider, D. & Shahabuddin, M. Malaria parasite development in a *Drosophila* model. *Science* **288**, 2376–2379 (2000).
61. Ribeiro, J. M. Blood-feeding arthropods: live syringes or invertebrate pharmacologists? *Infect. Agents Dis.* **4**, 143–152 (1995).
62. Bolshakov, V. N. *et al.* A comparative genomic analysis of two distant Diptera, the fruit fly, *Drosophila melanogaster*, and the malaria mosquito, *Anopheles gambiae*. *Genome Res.* **12**, 57–66 (2002).
63. Hoffman, S. L. Research (genomics) is crucial to attacking malaria. *Science* **290**, 1509 (2000).
64. Hoffman, S. L. *et al.* Protection of humans against malaria by immunization with radiation attenuated *Plasmodium* sp. sporozoites. *J. Infect. Dis.* (in the press).
65. Schofield, L. *et al.* Gamma interferon, CD8⁺ T cells and antibodies required for immunity to malaria sporozoites. *Nature* **330**, 664–666 (1987).
66. Weiss, W. R., Sedegah, M., Beaudoin, R. L., Miller, L. H. & Good, M. F. CD8⁺ T cells (cytotoxic/suppressors) are required for protection in mice immunized with malaria sporozoites. *Proc. Natl Acad. Sci. USA* **85**, 573–576 (1988).
67. Doolan, D. L. & Hoffman, S. L. The complexity of protective immunity against liver-stage malaria. *J. Immunol.* **165**, 1453–1462 (2000).
68. Sim, B. K. *et al.* Induction of biologically active antibodies in mice, rabbits, and monkeys by *Plasmodium falciparum* EBA-175 region II DNA vaccine. *Mol. Med.* **7**, 247–254 (2001).
69. Bouharoun-Tayoun, H., Oeuvray, C., Lunel, F. & Drullin, P. Mechanisms underlying the monocyte-mediated antibody-dependent killing of *Plasmodium falciparum* asexual blood stages. *J. Exp. Med.* **182**, 409–418 (1995).
70. Hoffman, S. L., Rogers, W. O., Carucci, D. J. & Venter, J. C. From genomics to vaccines: malaria as a model system. *Nature Med.* **4**, 1351–1353 (1998).
71. Collins, F. H., Kamau, L., Ranson, H. A. & Vulule, J. M. Molecular entomology and prospects for malaria control. *Bull. World Health Organ.* **78**, 1412–1423 (2000).
72. Hemingway, J. & Ranson, H. Insecticide resistance in insect vectors of human disease. *Annu. Rev. Entomol.* **45**, 371–391 (2000).
73. Gentile, G., Slotman, M., Ketmaier, V., Powell, J. R. & Caccone, A. Attempts to molecularly distinguish cryptic taxa in *Anopheles gambiae* s.s. *Insect Mol. Biol.* **10**, 25–32 (2001).
74. Enserink, M. Malaria research: two new steps toward a 'better mosquito'. *Science* **293**, 2370–2371 (2001).
75. Catteruccia, F. *et al.* Stable germline transformation of the malaria mosquito *Anopheles stephensi*. *Nature* **405**, 959–962 (2000).
76. Mills, C. D., Burgess, D. C., Taylor, H. J. & Kain, K. C. Evaluation of a rapid and inexpensive dipstick assay for the diagnosis of *Plasmodium falciparum* malaria. *Bull. World Health Organ.* **77**, 553–559 (1999).
77. Djimde, A. *et al.* A molecular marker for chloroquine-resistant falciparum malaria. *N. Engl. J. Med.* **344**, 257–263 (2001).
78. Kublin, J. G. *et al.* Molecular assays for surveillance of antifolate-resistant malaria. *Lancet* **351**, 1629–1630 (1998).
79. Fortin, A. *et al.* Identification of a new malaria susceptibility locus (Char4) in recombinant congenic strains of mice. *Proc. Natl Acad. Sci. USA* **98**, 10793–10798 (2001).
80. Roses, A. D. Pharmacogenetics and the practice of medicine. *Nature* **405**, 857–865 (2000).
81. Dervan, P. B. & Burli, R. W. Sequence-specific DNA recognition by polyamides. *Curr. Opin. Chem. Biol.* **3**, 688–693 (1999).
82. Menard, R. & Janse, C. Gene targeting in malaria parasites. *Methods* **13**, 148–157 (1997).
83. Mota, M. M., Thathy, V., Nussenzweig, R. S. & Nussenzweig, V. Gene targeting in the rodent malaria parasite *Plasmodium yoelii*. *Mol. Biochem. Parasitol.* **113**, 271–278 (2001).
84. Caffrey, C. R., Scory, S. & Steverding, D. Cysteine proteinases of trypanosome parasites: novel targets for chemotherapy. *Curr. Drug Targets.* **1**, 155–162 (2000).
85. Curtis, C. F. Infectious disease. The case for deemphasizing genomics in malaria control. *Science* **290**, 1508 (2000).

Acknowledgements

We are extremely grateful to A. Holser for expert bibliographic assistance in preparing this manuscript.

Satellite imagery in the study and forecast of malaria

David J. Rogers*, Sarah E. Randolph†, Robert W. Snow‡§ & Simon I. Hay*‡

*TALA Research Group and †Oxford Tick Research Group, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK (e-mail: david.rogers@zoo.ox.ac.uk)

‡Kenya Medical Research Institute/Wellcome Trust Collaborative Programme, PO Box 43640, Nairobi, Kenya

§Centre for Tropical Medicine, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DU, UK

More than 30 years ago, human beings looked back from the Moon to see the magnificent spectacle of Earth-rise. The technology that put us into space has since been used to assess the damage we are doing to our natural environment and is now being harnessed to monitor and predict diseases through space and time. Satellite sensor data promise the development of early-warning systems for diseases such as malaria, which kills between 1 and 2 million people each year.

Malaria caused by *Plasmodium falciparum* parasites exacts its greatest toll in sub-Saharan Africa, where it is one of the largest causes of morbidity and mortality, creating a significant barrier to economic development. Furthermore, this public health burden is increasing globally¹, exacerbated by failure of existing affordable drugs, population growth against declining *per capita* expenditure on health, human migration and poverty. As a step towards reversing this trend, there is growing interest in the mapping and predictive modelling of the geographical limits, intensity and dynamics of the risk of malaria infection, using new tools of surveillance. An unprecedented amount of information on environmental conditions, remotely sensed by satellite sensors, is now available at temporal and spatial resolutions to match our epidemiological questions.

Here we show how these tools are used to investigate the factors that drive the dynamics of vector populations and malaria parasite transmission. Because mosquito population processes and malaria incubation periods in vectors, for example, vary with temperature and moisture conditions on the ground, remotely sensed images of seasonal climate are

powerful predictors of mosquito distribution patterns and average levels of transmission of malaria parasites by these vectors. Patterns of infection vary through time owing to extrinsic (for example, climate) and intrinsic (for example, immunity) effects. The balance of these factors depends upon the levels of malaria transmission in each place and will change over time with resistance to control of parasites and vectors. Early-warning systems, therefore, will require models that incorporate both intrinsic and extrinsic factors.

Extrinsic and intrinsic drivers of malaria

Diseases caused by vector-borne pathogens vary in magnitude through space and time much more than directly transmitted pathogens, because their innate capacity to increase is usually much higher. This is expressed as the basic reproductive number R_0 :

$$R_0 = ma^2bce^{-\mu T}/\mu r$$

where m is the vector–host ratio; a is the vector biting rate; b, c are transmission coefficients from vertebrate to vector and vice versa; μ is the vector mortality rate; T is the extrinsic incubation period of the parasite in the vector; and r is the rate of recovery in the vertebrate. Values of R_0 for vector-borne pathogens reach hundreds, or even thousands, compared with a typical value of <10 for directly transmitted pathogens². Changes in R_0 are most sensitive to changes in variables that appear as powers or exponents (a, μ and T)^{3,4}.

R_0 is reduced to R_e , the effective reproductive number that approximates to 1.0 at equilibrium, by population acquired immunity and other processes that either limit vertebrate infectivity for mosquitoes or decrease coefficients of transmission between vertebrates and vectors. Thus, although the increase of vector-borne pathogens is most strongly influenced by the six (out of seven) vector-related factors in the R_0 equation, their regulation is effected principally by vertebrate factors. However, because acquired vertebrate immunity depends upon the level and frequency of exposure to parasites since birth, it should be possible to define the dynamics of many vector-borne pathogens solely on the basis of the current and previous history of the vector population.

Although invertebrate vectors are strongly influenced by variable climate (that is, abiotic or extrinsic factors)⁵, vertebrate host immunity effects (biotic or intrinsic factors)

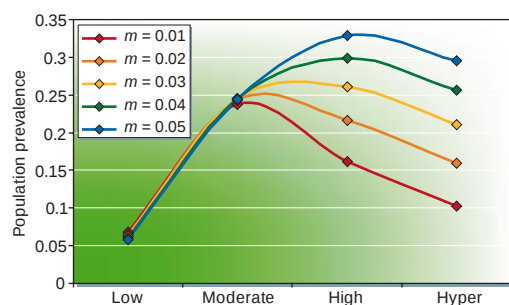
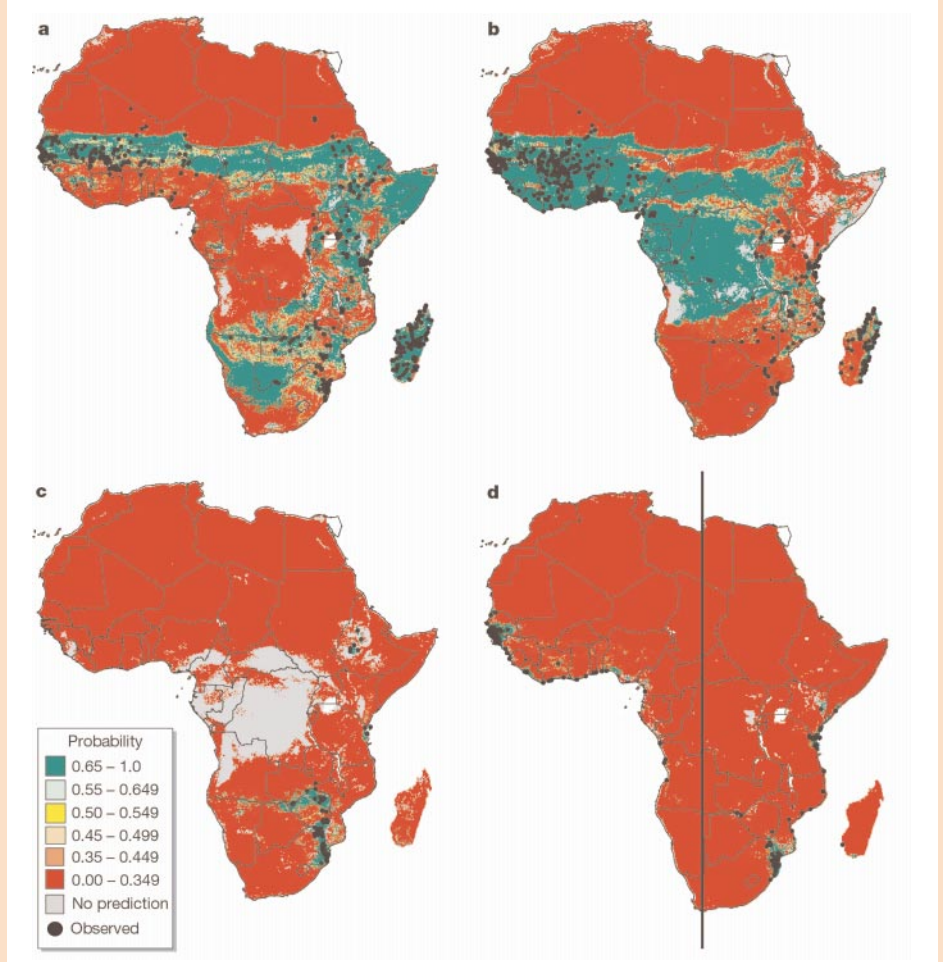


Figure 1 Hypothetical relationship between the challenge to a host population by a vector-borne pathogen and the risk of the host becoming infected. Challenge is a function of many elements of the vector's biology; risk is often modulated by host resistance and/or acquired immunity. Fixed age-specific prevalences at each level of challenge with different levels of vertebrate host mortality rate (m in the figure) produce population prevalence/challenge curves of different overall shapes.

Figure 2 Distributions of five mosquito species in the *Anopheles gambiae* complex in Africa, predicted from temporal Fourier-processed satellite data (Box 1) and elevation (global coverage provided by the digital elevation model GTOPO30; <http://edcdaac.usgs.gov/gtopo30/README.html>) at a spatial resolution of 0.05°. The colour-coded probabilities⁴² of presence effectively indicate the environmental suitability for each species throughout the continent. Symbols indicate sample sites for each species. Between 18° N to 30° S, each species was classed as present only within 0.15° of the sites from which it has been recorded²⁴ and absent only from similarly sized sites where any of the other species have been recorded. Within these sites, presence pixels (300 for *A. gambiae* s.s. and *A. arabiensis* and 100 for the other species) and absence pixels (400 for all species) were chosen at random; additional randomly selected absence pixels were chosen north of 18° N ($n=200$) and south of 30° S ($n=50$). Satellite data: middle infrared, land surface temperature and normalized difference vegetation index for 1982–1998 were derived from the National Oceanographic and Atmospheric Administration's Advanced Very High Resolution Radiometer, and cold cloud duration for 1988–1999 from Meteosat High Resolution Radiometer. Satellite and elevation data for all sample pixels were subjected to k -means clustering within the Statistics Package for the Social Sciences (SPSS, Chicago) to identify up to six natural clusters each of presence and absence pixels for each mosquito species. Within maximum likelihood discriminant analysis, stepwise selection of up to ten variables was applied to maximize predictive accuracy according to the kappa statistic²⁸, sensitivity and specificity (Box 1) and to calculate the posterior probabilities⁴² with which each pixel belongs to the presence or absence classes within the training set. Sites too different from any of the training set sites are assigned to a 'no prediction' class. **a**, *A. arabiensis*: 86.0% correct predictions, 7.7% false positives, 6.2% false negatives, sensitivity = 0.80, specificity = 0.89, $\kappa = 0.679$ (± 0.051 95% confidence interval). **b**, *A. gambiae* s.s.: 92.4% correct, 4.2% false positives,



3.4% false negatives, sensitivity = 0.89, specificity = 0.94, $\kappa = 0.826$ (± 0.039). **c**, *A. quadriannulatus*: 99.1% correct, 0.9% false positives, 0% false negatives, sensitivity = 1, specificity = 0.99, $\kappa = 0.961$ (± 0.029). **d**, *A. melas* (West Africa): 98.2% correct, 1.6% false positives, 0.1% false negatives, sensitivity = 0.99, specificity = 0.98, $\kappa = 0.928$ (± 0.039); and *A. merus* (East Africa): 98.4% correct, 1.6% false positives, 0% false negatives, sensitivity = 1, specificity = 0.98, $\kappa = 0.933$ (± 0.037).

can produce characteristic cycles of infection even in the absence of variation in environmental conditions⁶. The interaction of extrinsic and intrinsic factors results in the characteristic waxing and waning of many vector-borne infections, on timescales from weeks to decades^{7,8}; these cycles may be suppressed, but are rarely totally eliminated, by control and intervention efforts. Early attempts to predict malaria outbreaks in India and Pakistan revealed an understanding of this interaction⁹: a combination of the absence of malaria in the preceding 5 years, rainfall anomalies from July to August, and the local prices of wheat (thought to reflect the nutritional state of the human population) was used to predict the amount and distribution of malaria in 1921 with 'considerable precision'¹⁰. This early promise of accurate forecasting of malaria outbreaks was implemented successfully in the subcontinent for almost 25 years¹¹ until the introduction of insecticides and cheap drugs in the 1950s and 1960s, when forecasting was no longer felt necessary¹². There have been no effective early-warning systems for malaria, or indeed for any vector-borne diseases, since that time¹³.

New prospects for malaria early-warning systems

With a breakdown in effectiveness of the 'cures' found for malaria in the 1950s and 1960s, there has been a global resurgence of

malaria^{1,14,15} and renewed interest in the concept of malaria early warning^{13,16,17}. But instead of rain gauges and wheat prices, we now have an armoury of information on environmental conditions, remotely sensed from satellite sensors, that has been related to the dynamics of vector populations and pathogen transmission^{18,19}. Until we can dissect quantitatively the roles played by extrinsic and intrinsic factors, however, we cannot use these new tools to forecast outbreaks. First, therefore, we focus on what satellite imagery can tell us about the environmental prerequisites for malaria transmission in the equilibrium situation.

The risk–challenge relationship for malaria

The hypothetical relationship between the challenge presented to the human population by a vector population and the resulting incidence or prevalence of infection, each assuming stable conditions, is complex and nonlinear (Fig. 1). The humped curve is determined by interactions between transmission rates, the rate of development and duration of temporary acquired immunity in the vertebrate host population, and the age structure of the latter. But the precise shape of the relation, and its implications for malaria control, are controversial^{20,21}, not least because of a shortage of good quality field data available for its determination. (The relationship between

clinical disease and challenge may be different from that shown in Fig. 1, most obviously in areas of high challenge.) The data that do exist for Africa, now being gathered together in the ambitious Mapping Malaria Risk in Africa/Atlas du Risque de Malaria en Afrique collaboration^{22,23}, may nevertheless be used first to examine the 'challenge' axis of Fig. 1.

Distribution and abundance of mosquito vectors

The distribution of efficiently transmitted pathogens such as those that cause malaria is generally limited by the distribution of competent vectors, which can now be predicted from satellite imagery. In Africa, the main vectors of malaria include species of the *Anopheles gambiae* complex (*A. arabiensis*, *A. bwambae*, *A. gambiae* s.s., *A. melas*, *A. merus* and *A. quadriannulatus*) and *A. funestus*, whose distributions show similarities with patterns of annual rainfall across Africa²⁴. Furthermore, a relationship found in West Africa between climate (annual precipitation and annual and wet-season temperatures) and the ratio of *A. gambiae* s.s. to *A. arabiensis* was used successfully to predict the distribution and relative abundance of these two species in Tanzania (East Africa)²⁵. The subtlety and extensive coverage of multivariate climatic factors detectable by satellite sensors have allowed predictions of the distribution of five of the six species in this complex (Fig. 2; *A. bwambae*, restricted to a small site in Uganda, is not modelled). This has been achieved using maximum likelihood methods based on the relatively few studies that identified these species separately. Thus, satellites can distinguish between the habitats of species that were, until 1956, regarded as a single species.

Satellite data show strong relationships with the density of other vectors in Africa^{2,26}, but mosquito abundance has only rarely been recorded, despite its importance in R_0 calculations. Until such data are collected, as a matter of urgency, we speculate that abundance will

be greatest in sites with conditions near to the climatic centroids defined by the satellite data, although spatially variable, density-dependent factors may confound this prediction.

Entomological inoculation rate

A more complete measure of malaria challenge is an estimate of the entomological inoculation rate (EIR), which is the product of the number of mosquito bites per human per unit time (ma in the R_0 equation) and the proportion of mosquitoes with sporozoites (infective stages of the malaria parasites)³. EIRs are generally derived from short- to medium-term studies, each in a relatively small area, and are therefore of limited predictive value on their own²⁷. By compiling all reliable published EIRs and relating them to satellite data, we can make more extensive predictions of this measure of challenge (Fig. 3). Analysis shows that gradual changes in environmental conditions distinguish low- and high-risk areas, and do so consistently enough for an excellent overall agreement with the field data (kappa statistic (ref. 28; and see Box 1) of 0.77). There is also good agreement between the EIR predictions (Fig. 3) and the predicted distributions of the two key malaria vectors, *A. arabiensis* and *A. gambiae* s.s. (Fig. 2), although in parts of southern Africa, where historically vector control and case-management have been effective, present-day malaria challenge is less than is shown on this map. But many areas of Africa are too different from any of the training sets to allow any predictions for them (shown in grey in Fig. 3). By identifying these regions, satellite imagery can direct new studies of this important malariometric index.

Resulting malaria incidence and prevalence

Biologically, and for the purposes of intervention, the point of interest is to explain the relationship between the above three measures of challenge (mosquito distribution and abundance, and EIR) and the resulting disease burdens, and so be able to predict the latter. We are hindered from making even statistical predictions by the lack of good quality training sets for satellite studies, as disease risks have been determined too infrequently, and over insufficiently wide areas. With the aid of satellites, however, a small training set can produce predictions across a large area.

Numbers of monthly childhood malaria admissions in three hospitals in Kenya, expressed as a percentage of the annual totals, were found to be correlated most consistently (mean adjusted $r^2 = 0.71$) with the previous month's normalized difference vegetation index (NDVI), (Box 1)²⁹, which is related to plant photosynthetic activity. As minimum NDVIs of 0.35–0.40 were shown to be required before >5% of annual admissions were recorded in any one month, the duration of malaria transmission seasons across Kenya and Uganda could be predicted by counting the number of months in which these NDVI values were exceeded^{18,30}. The resulting predictive maps of malaria seasons showed strong similarities to an historical map of malaria transmission periods in Kenya³¹.

Non-equilibrium situations

Seasonality and intrinsic dynamics

The position of any site on the horizontal axis of Fig. 1 determines average malaria risk, but seasonal changes in the challenge variables (mosquito abundance and EIR) cause seasonal changes in risk that may be either very pronounced (such as is characteristic of conditions of low average challenge) or strongly buffered by herd immunity responses (at high challenges). At low challenge levels, and because of time delays between the index of challenge and its expression as malaria infections, the risk–challenge relationship will trace an anticlockwise ellipse over time, completing approximately one cycle per year. Remotely sensed data that monitor one or more measure of challenge will therefore be able to predict risk with an appropriate lead time.

In high challenge areas this simple picture is obscured by the longer-term effects of intrinsic disease dynamics, which tend to be uncoupled from the annual seasonality in a way that makes

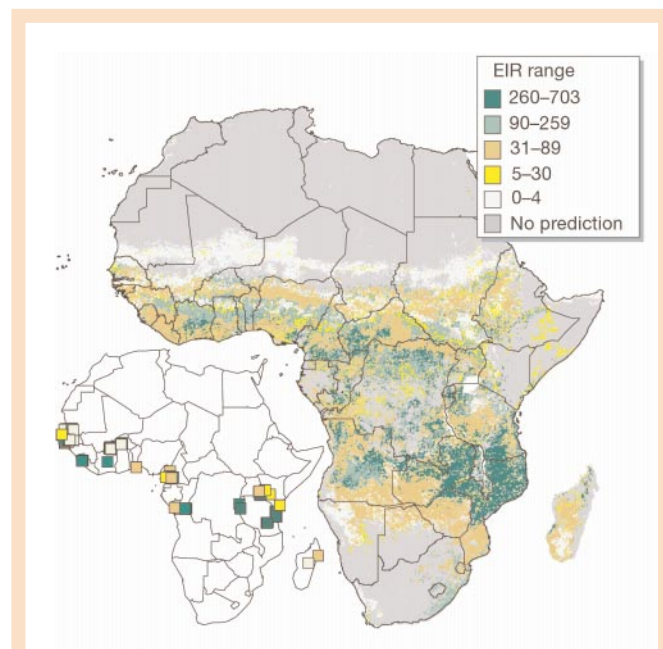


Figure 3 Satellite-derived predictions of entomological inoculation rate (EIR) in Africa. EIR data²⁷ (map inset) were grouped into five approximately equal-sized classes of mean levels of malaria challenge. The same satellite data layers and analytical methods as used in Fig. 2 are used to define the probability with which each continental pixel belongs to one of the five challenge categories. Insufficient training data were available to define EIR in those parts of the continent marked grey. Descriptive accuracies: EIR range 0–4.4, 85.7% ($n = 21$); 5.8–26.6, 81.8% ($n = 22$); 31.0–87.0, 77.3% ($n = 22$); 89.8–255.5, 68.2% ($n = 22$); 259.9–703.4, 95.5% ($n = 22$). Overall $\kappa = 0.771 (\pm 0.064)$.

Box 1

Satellite sensors for monitoring diseases

Satellite sensor designs are rarely ideal for epidemiological studies because of trade-offs between spectral, spatial and temporal resolution, determined by constraints of the Earth's atmosphere, or the original requirements of commissioning agencies. Passive satellite sensor data (that is, reflections or emissions arising ultimately from the Sun) have been used most commonly for epidemiological studies, and are discussed here, but there is increasing interest in radar satellites with active sensors that can produce images even under cloudy conditions.

Spectral resolution

Satellite sensors detect reflected sunlight or infrared radiation emitted by all bodies above absolute zero. Data are most readily available in three to seven wavebands or channels in the human-visible and near-to-thermal infrared part of the electromagnetic spectrum (0.3–14- μ m wavelengths).

Spatial resolution

Earth-observing satellites produce data with spatial resolutions of 1–4 m (Ikonos-2), 10–20 m (Satellite pour l'Observation de la Terre; SPOT), 30–120 m (Landsats 1-5) or 15–60 m (Landsat 7). Images, made up of picture elements or 'pixels' of these sizes, have swath widths of ~11 km (Ikonos), ~60 km (SPOT) and 185 km (Landsat). The 'vegetation instrument' on SPOT-4 has a spatial resolution of 1 km and a 2,250-km swath width.

Meteorological satellites have lower spatial resolutions, with pixel sizes down to 1.1 km (National Oceanographic and Atmospheric Administration Advanced Very High Resolution Radiometer; NOAA-AVHRR), and a correspondingly wider swath width of ~2,400 km. Geostationary satellites maintain a constant position relative to the Earth, giving spatial resolutions of 1–8 km (Geostationary Operational Environmental Satellite for the Americas) or 2.5–5 km (Meteosat 4–6 for Europe/Africa) and images of the entire Earth half-disk.

Temporal resolution

Satellites with a higher spatial resolution have a repeat frequency of 11 (Ikonos), 16 (Landsat) or 26 (SPOT) days. Orbiting meteorological satellites produce two images per day of the entire Earth's surface, whereas geostationary satellites produce two images per hour to monitor weather systems.

New satellites and sensors

New systems promise greater spectral and spatial resolutions, and greater signal stability over time. These include the Moderate Resolution Imaging Spectrometer on the Terra spacecraft (<http://terra.nasa.gov>), and the Spinning Enhanced Visible and Infrared Imager on the geostationary Meteosat Second Generation (<http://www.esa.int>).

Images for epidemiology

Imagery is adversely affected by atmospheric contamination such as clouds and other aerosols. The low repeat frequency of the higher spatial resolution satellites prevents the recording of important seasonal determinants of pathogen transmission rates. In contrast, frequent images from NOAA-AVHRR and Meteosat sensors can be combined to produce relatively cloud-free monthly images generated from the maximum values of each signal recorded during the period (assumed to reflect cloud-free conditions), called maximum value composites¹⁸. These are of much greater use in studying dynamical epidemiological processes.

Data from each satellite channel may be used directly to describe epidemiological events, or may be processed to produce indices related to ground-based variables such as soil surface temperatures. Commonly used products include the middle infrared band, derived from AVHRR channel 3, and land surface temperature (LST), derived from AVHRR channels 4 and 5, both of which are related to the Earth's surface temperature; the normalized difference vegetation index, derived from AVHRR channels 1 and 2 and related to plant photosynthetic activity; near-surface air temperature, derived from LST and vegetation index measurements; and cold cloud duration from Meteosat, which is correlated with rainfall in convective precipitation systems (all reviewed in refs 18,45).

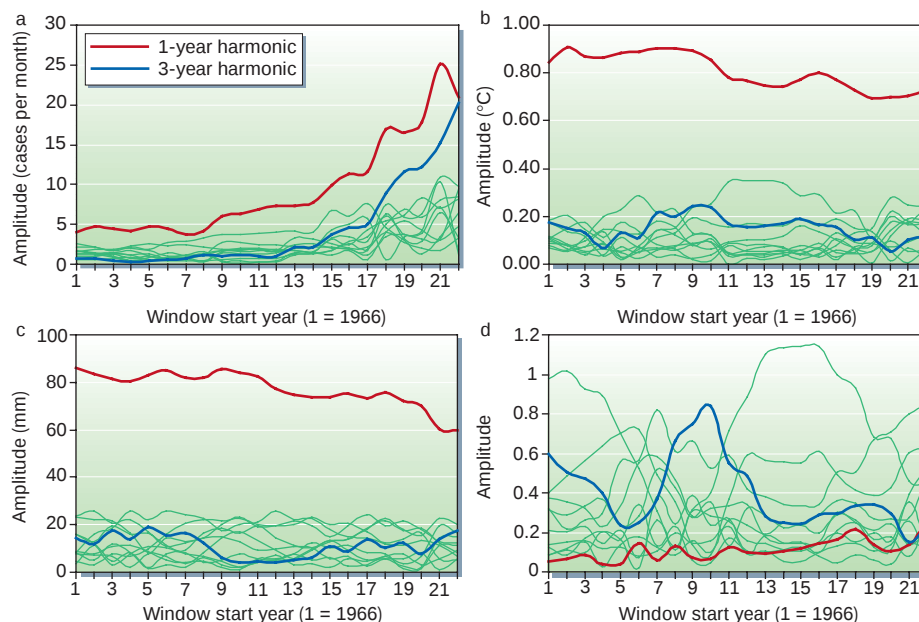
Data processing and application

Monthly composited imagery often shows strong serial correlations, and therefore data redundancy, which may be overcome in two ways. The data may be subjected either to principal components analysis (PCA) and the resultant significant principal components used in further analyses, or to temporal Fourier analysis that describes natural cycles in terms of annual, bi-annual, tri-annual, and so on, components with longer or shorter periods. Temporal Fourier processing removes data redundancy and produces a set of orthogonal (uncorrelated) outputs while retaining a description of seasonality (which is lost in PCA); this is of vital interest in epidemiology^{26,46,47}. Windowed Fourier analysis overcomes the problem of serial changes in the mean and variance of the data. Trends are first removed by taking the difference of the time series from a moving average spanning a number of annual cycles, and this de-trended time series is then Fourier-analysed. To deal with changing variances, a fixed aperture window is moved progressively over the de-trended time series, and only the data within the window are analysed. By comparing such analyses across the entire time series, the periodicities contributing most to changes in the overall variance can be identified (Fig. 4).

Composited, multi-temporal and Fourier-processed satellite sensor data may be used to describe epidemiological data statistically using linear and logistic regression techniques^{48,49} or various discriminant analysis and maximum likelihood approaches^{26,50}. Output maps record the similarity of each pixel to the satellite-determined environmental characteristics in a sample set of sites where the epidemiological situation is well documented (the training set). Such maps can record the predicted suitability of each pixel for the presence of a vector or disease (Fig. 2), or quantitative data related to the burden of disease (Fig. 3).

Predictive accuracy can be assessed using a contingency table that compares the training set data and the suitability category to which the pixels were assigned. From this is calculated the overall percentage of correct predictions, the percentage of false positives and false negatives (that is, false predictions of presence and absence, respectively), and the sensitivity and specificity (proportion of positives or negatives, respectively, correctly identified). The kappa index of agreement, κ , measures predictive accuracy compared with a null model (that is, one with no predictive skill); values vary between 0 (fit no better than random) and 1.0 (perfect fit)²⁸, with a value of more than 0.75 regarded as excellent; confidence intervals can be attached to κ values. Once robust and reliable correlations between the satellite and disease data are established, real-time monitoring of environmental conditions by satellites can provide valuable inputs into disease early-warning systems¹⁶.

Figure 4 Amplitude of Fourier harmonics derived from windowed Fourier analysis of malaria cases per month and a range of climatic variables for the period January 1966 to December 1998. **a**, Malaria cases per month; **b**, temperature (°C); **c**, rainfall (mm); **d**, Data from the Multivariate El Niño/Southern Oscillation (ENSO) Index^{43,44} (<http://www.cdc.noaa.gov/~kew/MEI/>). All data were first de-trended with a 60-point moving average, and the analysis was based upon the difference of the raw data from this trend line. Beginning in 1966, a 12-year window was moved forward one year at a time until 1998 was included. The rapid increase in the amplitudes of the 1-year (red line) and ~3-year (blue line) harmonics of the malaria data have no obvious parallels in the temperature, rainfall or ENSO records, suggesting that these changes are not driven by any element of climate change. The green lines show the results for all other harmonics with periods of between 1 and 12 years.



prediction much less accurate. Intrinsic dynamics can produce cycles of disease prevalence in human populations with periods of a few to many years. A recent analysis of long-term records of two mosquito-borne diseases, dengue in Thailand and malaria in Kenya, shows clear evidence of such cycles, with periods in each case of about 3 years³². The absence of any sign of equivalent multi-annual cycles in the contemporary meteorological records at each site supports the interpretation that these cycles are driven by intrinsic biotic factors such as host acquired immunity.

The transition from low levels of endemicity, where variation in malaria incidence is apparently driven by extrinsic factors, to higher levels of endemicity, where intrinsic dynamics overrides annual seasonality, may be abrupt (D.J.R., unpublished data). This transition may occur with only modest increases in average mosquito abundance. Above a certain level of endemicity the resulting multi-annual cycles of infection are neither simply, nor easily, linked to either meteorological or satellite data.

Short- and long-term cycles in weather patterns

Longer-term weather cycles such as the El Niño/Southern Oscillation (ENSO) in the Pacific Ocean have been invoked recently to 'explain' outbreaks of malaria and other diseases³³. Some of these analyses are not statistically robust, while none of them allows an alternative explanation involving intrinsic cycles. This problem is particularly acute because within the past 30–40 years the quasiperiodic ENSO signal has shown a gradual increase in frequency and now has a periodicity of approximately four years³⁴. This is close enough to the intrinsic cycles revealed for dengue in Thailand and malaria in Kenya for a casual analysis to suggest a causal link between the two. Before any disease periodicity can be attributed with any certainty to ENSO, a link must be established between the global ENSO index and local weather records (measured directly or remotely) pertinent to the particular vector-borne pathogen.

Trends and global change

The disparity in conclusions about the likely impact of global climate change on malaria distribution in the future^{35,36} highlights the significant gaps in our knowledge of this most important tropical disease. Attempts to relate past trends in malaria prevalence to climate records³⁷ are the critical tests of the real role of climate in the recent

history of malaria and therefore its likely impact in the future. Statistical analyses of time series usually require stationarity (the property of constant mean and variance through a series), which is generally achieved by calculating the difference from the overall trend line of each sequential observation. Often, however, it is the trend itself which is of interest and importance.

Windowed Fourier analysis (Box 1) of long-term (1966–1998) monthly malaria data from Kericho, Kenya^{32,38}, shows that the amplitude of many of the component Fourier harmonics has increased over this period; this is most marked in cycles with periods of one and about three years (Fig. 4). Similar analyses of contemporary temperature and rainfall records show no equivalent changes (Fig. 4); rather, if anything there has been a decrease in the amplitude of the same Fourier harmonics. The most parsimonious interpretation of these results is that following a period of aggressive and wide-spread use of chloroquine for fever management, and for some periods as prophylaxis, this drug became less effective at parasite clearance during the mid–late 1980s. Malaria no longer suppressed by the drugs began to show higher-amplitude extrinsic and intrinsic cycles, coupled, respectively, to annual seasonal variation and infection–recovery–immunity effects.

Where climate change is accompanied by changing abundance of mosquitoes, its impact on the levels and periodicity of malaria outbreaks will be far more complex than modelled so far. Although remotely sensed images of seasonal environments can set the scene for long-term studies of disease, such data must be combined with contemporary field data and dynamic models of disease transmission for a full explanation of vector-borne diseases in an ever-changing world.

Future perspectives

Statistical models relating various components of the transmission of malaria parasites to satellite data demonstrate clearly the potential role and importance of remotely sensed imagery to descriptions, explanations and predictions of vector-borne disease. In the face of a world changing in both abiotic and biotic respects, and complex infection and disease dynamics that are not amenable to simple statistical interpretations, we must now take the next step — the incorporation of satellite data into biological models of pathogen transmission. Studies have already related vector mortality rates and

abundance to satellite data³⁹, and biological models have been developed for a few vectors^{40,41} and vector-borne pathogens²⁶.

Extensive satellite coverage coupled with vector-borne pathogen models that are appropriate at a local scale will enable us to build spatially rich, accurate models of vector-borne pathogens. If we can understand transmission dynamics well enough to model the present, we should be able to develop accurate disease early-warning systems in the future. It is clear that the technologies we now have to study these diseases are far better than those available to malariologists in the early years of the last century. The challenge is to make the science of malaria prediction at least as good. □

1. Snow, R. W., Trape, J.-F. & Marsh, K. Childhood malaria mortality in Africa: past, present and future. *Trends Parasitol.* **17**, 593–597 (2001).
2. Rogers, D. J. Satellite imagery, tsetse and trypanosomiasis in Africa. *Prev. Vet. Med.* **11**, 201–220 (1991).
3. Macdonald, G. *The Epidemiology and Control of Malaria* (Oxford Univ. Press, London, 1957).
4. Onori, E. & Grab, B. Indicators for the forecasting of malaria epidemics. *Bull. World Health Organ.* **58**, 91–98 (1980).
5. Hay, S. I., Tucker, C. J., Rogers, D. J. & Packer, M. J. Remotely sensed surrogates of meteorological data for the study of the distribution and abundance of arthropod vectors of disease. *Ann. Trop. Med. Parasitol.* **90**, 1–19 (1996).
6. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, Oxford, 1991).
7. Jacob, K. B. M. & Swaroop, S. Investigation of long-term periodicity in the incidence of epidemic malaria in the Punjab. *J. Malaria Instit. India* **6**, 39–51 (1945).
8. Rogers, D. J. The dynamics of vector-transmitted diseases in human communities. *Phil. Trans. R. Soc. Lond. B* **321**, 513–539 (1988).
9. Christophers, S. R. Epidemic malaria of the Punjab, with a note on a method of predicting epidemic years. *Paludism* **2**, 17–26 (1911).
10. Gill, C. A. The prediction of malaria epidemics. *Ind. J. Med. Res.* **10**, 1136–1143 (1923).
11. Swaroop, S. Forecasting of epidemic malaria in the Punjab, India. *Am. J. Trop. Med.* **29**, 1–17 (1949).
12. Najera, J. A., Kouznetsov, R. L. & Delacollette, C. *Malaria Epidemics: Detection and Control Forecasting and Prevention* (World Health Organization, Geneva, 1998).
13. National Research Council. *Under the Weather: Climate, Ecosystems, and Infectious Disease* (Committee on Climate, Ecosystems, Infectious Diseases, and Human Health, Board on Atmospheric Sciences and Climate, National Research Council, Washington DC, 2001).
14. Snow, R. W., Craig, M., Deichmann, U. & Marsh, K. Estimating mortality, morbidity and disability due to malaria among Africa's non-pregnant population. *Bull. World Health Organ.* **77**, 624–640 (1999).
15. World Health Organization. *The World Health Report 1999: Making a Difference* (World Health Organization, Geneva, 1999).
16. Myers, M. F., Rogers, D. J., Cox, J., Flauhait, A. & Hay, S. I. Forecasting disease risk for increased epidemic preparedness in public health. *Adv. Parasitol.* **47**, 309–330 (2000).
17. World Health Organization. *Malaria Early Warning Systems, a Framework for Field Research in Africa: Concepts, Indicators and Partners* (World Health Organization, Geneva, 2001).
18. Hay, S. I. An overview of remote sensing and geodesy for epidemiology and public health application. *Adv. Parasitol.* **47**, 1–35 (2000).
19. Hay, S. I., Omumbo, J., Craig, M. & Snow, R. W. Earth observation, geographic information systems and *Plasmodium falciparum* malaria in sub-Saharan Africa. *Adv. Parasitol.* **47**, 173–215 (2000).
20. Snow, R. W. & Marsh, K. Will reducing *Plasmodium falciparum* transmission alter malaria mortality among African children? *Parasitol. Today* **11**, 188–190 (1995).
21. Smith, T. A., Leuenberger, R. & Lengeler, C. Child mortality and malaria transmission intensity in Africa. *Trends Parasitol.* **17**, 145–149 (2001).
22. Mapping Malaria Risk in Africa/Atlas du Risque de Malaria en Afrique (MARA/ARMA) collaboration. *Towards an Atlas of Malaria Risk in Africa. First Technical Report of the MARA/ARMA (Mapping Malaria Risk in Africa) Collaboration* (MARA/ARMA, Durban, 1998).
23. Craig, M. H., Snow, R. W. & le Sueur, D. A climate-based distribution model of malaria transmission in sub-Saharan Africa. *Parasitol. Today* **15**, 105–111 (1999).
24. Coetzee, M., Craig, M. H. & le Sueur, D. Distribution of African malaria mosquitoes belonging to the *Anopheles gambiae* complex. *Parasitol. Today* **16**, 74–77 (2000).
25. Lindsay, S. W., Parson, L. & Thomas, C. J. Mapping the ranges and relative abundance of the two principal African malaria vectors, *Anopheles gambiae sensu stricto* and *An. arabiensis*, using climate data. *Proc. R. Soc. Lond. B* **265**, 847–854 (1998).
26. Rogers, D. J. Satellites, space, time and the African trypanosomiasis. *Adv. Parasitol.* **47**, 129–171 (2000).
27. Hay, S. I., Rogers, D. J., Toomer, J. F. & Snow, R. W. Annual *Plasmodium falciparum* entomological inoculation rates (EIR) across Africa: literature survey, internet access and review. *Trans. R. Soc. Trop. Med. Hyg.* **94**, 113–127 (2000).
28. Congalton, R. G. A review of assessing the accuracy of classifications of remotely sensed data. *Rem. Sens. Environ.* **37**, 35–46 (1991).
29. Hay, S. I., Snow, R. W. & Rogers, D. J. Predicting malaria seasons in Kenya using multitemporal meteorological satellite sensor data. *Trans. R. Soc. Trop. Med. Hyg.* **92**, 12–20 (1998).
30. Hay, S. I., Snow, R. W. & Rogers, D. J. From predicting mosquito habitat to malaria seasons using remotely sensed data: practice, problems and perspectives. *Parasitol. Today* **14**, 306–313 (1998).
31. Butler, R. J. *Atlas of Kenya: A Comprehensive Series of New and Authenticated Maps Prepared from the National Survey and other Governmental Sources with Gazetteer and Notes on Pronunciation and Spelling* (The Survey of Kenya, Nairobi, 1959).
32. Hay, S. I. *et al.* Etiology of interepidemic periods of mosquito-borne disease. *Proc. Natl Acad. Sci. USA* **97**, 9335–9339 (2000).
33. Kovats, R. S., Bouma, M. J. & Haines, A. *El Niño and Health* (World Health Organization, Geneva, 1999).
34. McGregor, G. R. & Nieuwolt, S. *Tropical Climatology* (Wiley, Chichester, 1998).
35. Martens, W. J. M. *et al.* Climate change and future populations at risk of malaria. *Global Environ. Change* **9**, S89–S107 (1999).
36. Rogers, D. J. & Randolph, S. E. The global spread of malaria in a future, warmer world. *Science* **289**, 1763–1766 (2000).
37. Hay, S. I. *et al.* Climate change and the resurgence of malaria in the East African highlands. *Nature* (in the press).
38. Shanks, G. D., Biomndo, K., Hay, S. I. & Snow, R. W. Changing patterns of clinical malaria since 1965 among a tea estate population located in the Kenyan highlands. *Trans. R. Soc. Trop. Med. Hyg.* **94**, 253–255 (2000).
39. Rogers, D. J. & Randolph, S. E. Mortality rates and population density of tsetse flies correlated with satellite imagery. *Nature* **351**, 739–741 (1991).
40. Rogers, D. J. A general model for tsetse populations. *Insect Sci. Applic.* **11**, 331–346 (1990).
41. Randolph, S. E. & Rogers, D. J. A generic population model for the Africa tick *Rhipicephalus appendiculatus*. *Parasitology* **115**, 265–279 (1997).
42. Green, P. E. *Analyzing Multivariate Data* (The Dryden Press, Hinsdale, IL, 1978).
43. Wolter, K. The Southern Oscillation in surface circulation and climate over the tropical Atlantic, Eastern Pacific, and Indian Ocean as captured by cluster analysis. *Clim. Appl. Meteorol.* **26**, 540–558 (1987).
44. Wolter, K. & Timlin, M. S. Measuring the strength of ENSO—how does 1997/98 rank? *Weather* **53**, 315–324 (1998).
45. Goetz, S., Prince, S. & Small, J. Advances in satellite remote sensing of environmental variables for epidemiological applications. *Adv. Parasitol.* **47**, 289–307 (2000).
46. Rogers, D. J. & Williams, B. G. in *Large-Scale Ecology and Conservation Biology* (35th Symp. Br. Ecol. Soc./Soc. Conserv. Biol., Univ. Southampton, 1993) (eds Edwards, P. J., May, R. M. & Webb, N. R.) 249–273 (Blackwell Scientific Publications, Oxford, 1994).
47. Rogers, D. J., Hay, S. I. & Packer, M. J. Predicting the distribution of tsetse flies in West Africa using temporal Fourier processed meteorological satellite data. *Ann. Trop. Med. Parasitol.* **90**, 225–241 (1996).
48. Augustin, N. H., Muggleston, M. A. & Buckland, S. T. An autologistic model for the spatial distribution of wildlife. *J. Appl. Ecol.* **33**, 339–347 (1996).
49. Manel, S., Dias, J. M., Buckton, S. T. & Ormerod, S. J. Alternative methods for predicting species distributions: an illustration with Himalayan river birds. *J. Appl. Ecol.* **36**, 734–747 (1999).
50. Rogers, D. J. & Randolph, S. E. Distribution of tsetse and ticks in Africa: past, present and future. *Parasitol. Today* **9**, 266–271 (1993).

Acknowledgements

S.E.R. is currently supported by a NERC Senior Research Fellowship. R.W.S. is supported as a Senior Research Fellow by the Wellcome Trust. S.I.H. is currently supported as an Advanced Training Fellow by the Wellcome Trust. We thank M. Coetzee for supplying geo-referenced observations on the African distribution of the *A. gambiae* complex and D. Shanks for providing malaria incidence and meteorological data from the Brooke Bond Kericho Tea Estate.

The Medicines for Malaria Venture

Discovering developing and
delivering new and
affordable medicines for
this deadly disease

Medical science has made some dramatic breakthroughs in recent decades, which have the potential to reduce the global burden of death and disability from a wide variety of diseases. However, realizing this potential has, up to now, largely been restricted to diseases that can provide financial returns to companies willing to take on the costs of discovering and developing new treatments. Although market mechanisms generally work well to drive medical innovation, they can leave some global diseases of the poor, such as malaria, neglected.



Visit www.mmv.org
for more detail
on our R&D portfolio

Malaria remains endemic in more than 90 countries, principally in the developing world, with between 300 and 500 million new infections each year and 2.4 billion people vulnerable to infection. As well as causing more than one million deaths every year, mainly in children and pregnant women, morbidity associated with malaria also worsens poverty in affected countries, and those people who cannot afford \$30 or so for the most effective medicines will be regularly incapacitated with bouts of malaria fever. The cycle of ill health and poverty, and the compounding lack of market incentive for R&D, has been documented exhaustively in a recent report by the World Health Organization Commission on Macroeconomics and Health.¹

The need for affordable treatments

One of the key challenges in the fight against malaria (or any disease that affects the developing world) is not just to develop effective and safe treatments, but also to make sure they are available to local governments and people at a price that will allow widespread use. New anti-malarials are also needed because resistance has rapidly been building up against existing treatments.

The mission of the Medicines for Malaria Venture (MMV) is to renew and sustain the supply of affordable anti-malarials

through the discovery, development and delivery of effective new treatments for poor countries.

Public-private partnerships

In delivering its mission, MMV is using a fundamentally new approach to anti-malarial drug discovery and development, bringing the public and private sectors together to work on its projects.

Since its formation, MMV has forged a unique series of collaborations with pharmaceutical companies, non-governmental organizations, philanthropic organizations and academic institutions. MMV's portfolio is constantly growing and includes a wide range of projects to address the various medical needs in the treatment of malaria. We are delighted that four projects in our current portfolio are being carried out in partnership with GlaxoSmithKline, our co-sponsors in this Nature Insight.

For as long as malaria remains a public health problem, MMV will continue to bring different groups together to discover, develop and deliver effective and affordable drugs wherever they are needed most.

1. The Commission on Macroeconomics and Health. Macroeconomics and health: investing in health for economic development (World Health Organization, Geneva, 2001).

Christopher C. Hentschel, PhD
Chief Executive Officer, MMV

Medicines for Malaria Venture

Discover, Develop, Deliver

GlaxoSmithKline & malaria

Malaria is a public health problem in more than 90 countries and 40% of the global population is vulnerable to this potentially deadly disease. Around one million children die from malaria each year in Africa alone and resistance to the most commonly used and affordable anti-malarial drugs is developing rapidly.

A global health crisis needs to be tackled on a global scale, with the co-operation of governments, non-governmental organizations (NGOs) and private sector companies, as well as many other groups (see Figure 1). The fundamental problems of poverty and poor health systems cannot be addressed unless all sectors work together.

In fact, in recent years significant progress has been made to increase access to medicines in the developing world, and there is a growing sense of partnership between the different sectors involved. GlaxoSmithKline (GSK) remains at the forefront of this effort, and is the only company conducting research and development into the prevention and treatment of all three of the World Health Organization's (WHO's) top priority diseases – HIV/AIDS, TB and malaria.

Facing the Challenge

In June 2001 GSK published *Facing the Challenge*¹, a

document that outlines the contribution we make to improving healthcare in the developing world.

GSK's contributions include drug donation and preferential pricing programmes, the development of new medicines specifically for diseases prevalent in these areas, community partnerships with governments and NGOs, and a unique vaccine programme.

All these initiatives reflect the spirit of partnership needed to address the problems of malaria.

Preferential pricing

GSK has been offering substantial discounts on vaccines to governments, charities and agencies for public health programmes for nearly 20 years, and has been offering preferential prices to governments of developing countries for our HIV/AIDS anti-retroviral treatments since 1997. We believe that more can be done and so we have extended our preferential pricing offers to more products, countries and customers. This includes making anti-malarials available at sustainable preferential prices to all of the Least Developed Countries of the world and all of sub-Saharan Africa. GSK's commitments on preferential pricing are detailed in *Facing the Challenge*¹ and are summarized in Figure 2.

Vaccines

Global immunization is one of the most successful public health initiatives ever, and GSK leads the way in this field, as the world's largest vaccine manufacturer and one of the principal donors of vaccines. In conjunction with two other manufacturers, we have given 100 million doses of polio vaccine to African countries. In addition, as part of the Global Alliance for Vaccines and Immunization (GAVI), GSK's combination vaccine, *Tritanrix HB*TM, is being provided to African countries at a cost discounted by more than 90%.

No vaccine has yet been licensed to protect against malaria – the plasmodium parasites are highly complex, and are able to adapt quickly to the immune system's defences. However, in early 2001, GSK Biologicals and the Malaria Vaccine Initiative (MVI) joined forces to accelerate the development of a vaccine that has shown promising early results against malaria. Phase I trials of this compound have been underway in The Gambia for four years, and Phase II clinical trials have recently been expanded to children in Mozambique.

Other initiatives

In partnership with the World Bank, Oxford University and Lucknow University, we are providing millions of doses of albendazole for the largest randomized clinical trial ever undertaken. The project will involve 1.2 million children, and will assess the impact of early de-worming on long-term childhood survival.

Medicines for malaria

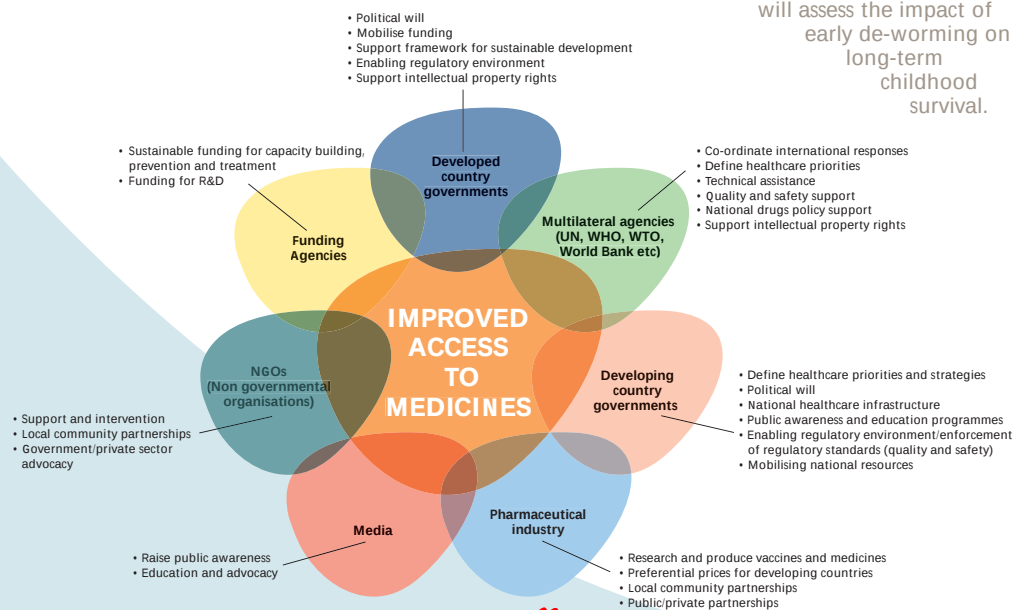
GSK already supplies 30 million malaria treatments per year in Africa at preferential prices, and we are now expanding this scheme to include our full range of anti-malarial medicines.

Since 1999, GSK has been involved in a groundbreaking collaboration to develop a new and affordable treatment for malaria. *Lapdap*TM (a combination of two existing compounds: chlorproguanil and dapsone) is effective against malaria in areas where other therapies have encountered resistance. It has been developed primarily for use in sub-Saharan Africa and will be priced at an affordable level. *Lapdap*TM is in the final stages of development, and regulatory files will be submitted in early 2002. The project is funded through a three-way partnership between GSK, the UK Department for International Development (DFID) and the WHO. The development of *Lapdap*TM has also involved medical teams from Liverpool University and the London School of Hygiene and Tropical Medicine, as well as collaborators across Africa.

GSK has recently built on the success of the *Lapdap*TM collaboration by becoming involved in two further projects. These are partially funded by the Medicines for Malaria Venture (MMV) and two new anti-malarial medicines are involved:

- Chlorproguanil/dapsone/artesunate (*Lapdap*TM plus artesunate) entered development in 2001. The combination of conventional anti-malarials with artemesin derivatives is increasingly seen as the preferred way of treating malaria while, at the same time, minimizing the risk of increasing drug resistance. The team that developed *Lapdap*TM is also carrying out the development of this triple combination, with funding from MMV, GSK, DFID and the WHO.
- Also in development is a novel analogue of amodiaquine, based on concepts developed at the University of Liverpool. Accelerated development of this compound, with MMV funding, is due to start in 2002.

Figure 1. Shared responsibilities





Malaria is also a risk to travellers visiting malaria endemic countries. Malarone™ (atovaquone and proguanil) is a novel combination medicine that is used both for prevention and as a treatment for uncomplicated malaria. It has been widely marketed in Europe and the USA. In addition, early-stage trials are exploring the use of a novel aminoquinolone, tafenoquine, for the prevention of malaria in travellers. This drug is currently in Phase III clinical trials.

At GSK, our mission is to improve the quality of human life by enabling people to do more, feel better and live longer. We know that no single organization can produce a solution to the crisis of malaria but, while we do not

have the mandate, expertise or resources to deliver healthcare unilaterally, we have taken a leadership role in addressing the issue of access to medicines. We believe that this is both an ethical imperative and the key to business success, and companies that respond sensitively to these challenges will be the leaders of the future. Our collaborations with MMV are one part of our commitment to these goals.

We are therefore proud to be associated with this Nature Insight, which will advance understanding of all aspects of the fight against malaria, the world's most deadly tropical disease.

GSK and diseases of the developing world

Drug donation programmes

Since 1998 GSK has played a leading role in the Global Alliance to Eliminate Lymphatic Filariasis (LF or elephantiasis – a major tropical parasitic disease causing lifelong disability in 80 countries). Working closely with the WHO, its regional offices and country Ministries of Health, GSK is an active and involved partner, providing disease-prevention treatments, funding and expertise on a massive scale. In 2001, GSK shipped 45 million treatments of its antiparasitic drug albendazole to 30 countries and provided about \$1 million in partner grants.

In 2002, it is estimated that 40 countries will request more than 100 million treatments as part of this scheme, and GSK is committed to continuing this long-term programme. Donations will increase to a maximum of 400–600 million treatments per year over 20 or more years – until LF is eliminated as a public health problem – making this potentially the largest drug donation programme in history. The WHO-recommended LF elimination strategy involves simple oral co-administration of two anti-parasitic drugs (albendazole with either diethylcarbamazine or ivermectin) given to entire at-risk endemic communities annually for 4–6 years.

Figure 2. GSK's preferential pricing commitments to the developing world

- Sustainable, preferential prices for all GSK HIV/AIDS medicines
- Anti-malarial medicines also offered at sustainable, preferential prices
- Expansion of countries eligible for preferential pricing offers. The offer will apply to all Least Developed Countries (as defined by the UN) and all countries of sub-Saharan Africa – a total of 63 countries
- Expansion of customer groups eligible for preferential pricing offers, including the Global Fund to Fight AIDS, TB and Malaria
- Development of a pilot programme to assess the impact of preferentially priced anti-infectives and de-worming agents in resource-poor settings.

These new initiatives build on the leadership role taken by GSK to address the barriers that block access to medicines in developing countries and complement the company's existing commitments of continued research and development into diseases of the developing world, offers of preferential prices on medicines and vaccines, and support for community activities that promote effective healthcare

Tritanrix HB™, Lapdap™ and Malarone™ are trademarks of the GlaxoSmithKline Group of Companies.
1 http://corp.gsk.com/community/facing_the_challenge.pdf



Contacts

Publisher: Fabien Savenay
Editor: Paul Smaglik
Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Group European Manager:

Leonie Welss (4954)

European Manager:

Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),
Andy Douglas (4975)

Holland/ Italy:

Nevin Bayoumi (4978)

Scandinavia:

Sille Opstrup (4994)

Spain/ Portugal:

Leonie Welss (4954)

Production Manager:

Billie Franklin

To send films and materials use

London address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau
Tel + 33 (0) 1 43 20 16 51
Fax + 33 (0) 1 43 20 51 52
e-mail: c.nioxchateau@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mails: p.phelan@nature.com
k.turner@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Director:

Ben Crowe

US Sales Manager:

Peyton Mason

US Advertising Coordinator:

Ashly de Leon

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Harakakatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746
e-mail: k.cowan@naturejp.com
Japan Manager: Kate Cowan

naturejobs

Swings and roundabouts

In the volatile world of pharmaceuticals and biotechnology, seeming paradoxes abound. Already, huge companies such as Glaxo Wellcome and SmithKline Beecham have merged, forming even bigger entities while simultaneously reorganizing themselves to act as if they were just a collective of small biotech firms.

Meanwhile, the more successful biotech companies are starting to resemble their pharmaceutical counterparts, not only in size but also in therapeutic approach. Genentech, for example, is considering dipping its toes into small-molecule therapeutics. And companies such as Celera, whose past president once said that it was primarily an information company, are now entering the race to find therapeutics.

The paradox — or, at least, confusion — continues when shifting the focus to employment (see pages 4–7). Many scientists switch company, or even form their own, around the time of a merger. Yet many mid-career academics are leaving tenured posts for pharmaceutical or biotech positions.

Perhaps this confusion stems from a more central paradox. Drug discovery, once the domain of blind screenings, is becoming more scientific in terms of target selection and lead-compound identification. Yet it is also becoming brutally industrial in terms of the speed and efficiency with which compounds can be screened. Job prospects in industry will depend on which of these factors is dominant. At the moment, science seems to be winning, which would help to explain the influx of academics into industry.

But the final paradox could yet occur. If drug discovery and development become too automated, scientists who were once attracted by the novelty could depart because of the resulting routine mundanity.

Paul Smaglik

Naturejobs editor



Contents

CAREERS AND RECRUITMENT

Keeping pace with drug discovery p4

The trend for outsourcing p7

WWW.NATUREJOBS.COM

Career centre
Information on the scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS

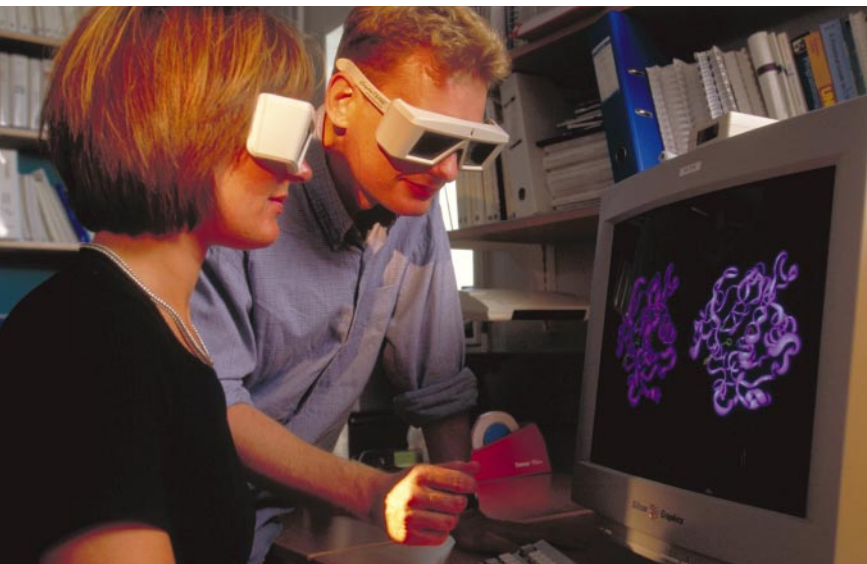
A route to flexible working

Researchers who aspire to work in drug discovery need to adapt to constantly changing technology and be able to harness new tools both to ask and to answer pertinent scientific questions, say Paul Smaglik and Adam Smith.

Structural biology was going to rationalize drug design. Next, combinatorial chemistry was to be the industrial panacea. More recently, the brute force of high-throughput screening bore the brunt of the buzz. But none of these, by themselves, has revolutionized drug discovery.

Technologies that drive these approaches, such as mass spectrometry and X-ray crystallography, are now being put in perspective — they are just part of an ever-evolving set of tools with which companies need to keep pace. This changing technological landscape is also affecting employment. Scientists who are conversant in many technologies, and flexible enough to learn new ones, are increasingly in demand. And those who can also step back, ask broad questions and then decide which technologies to use to answer them are at a premium.

Technology transfer: the rapidly changing face of drug discovery means researchers need to be versatile.



Companies are now diverting more of their resources to evaluate other companies' drug-development technologies, and they are placing more emphasis on recruiting scientific employees with broad backgrounds. Technology is also making some foundation skills less marketable, while creating demands for other skills that have not traditionally been associated with drug development. For the skills in shortest supply, companies may sponsor studentships for scientists who show both promise and interest in drug discovery.

PROBLEM-SOLVERS WANTED

This means that both pharmaceutical and biotech firms are less willing to commit to a specific skill set that could become obsolete. "We prefer broadly trained people who can move with the times," says

Yvonne Martin, head of computer-assisted molecular design at Abbott Laboratories, a drug company based in Illinois.

Such an approach can provide career longevity. "If you recruit someone for a particular skill, it may be useful for the first six months of their career, but it's not particularly useful for the next nineteen-and-a-half years," says Anthony Ford-Hutchinson, an executive vice-president at Merck Research Laboratories in West Point, Pennsylvania.



Anthony Ford-Hutchinson wants to recruit scientists with a broad base of skills.

This generalist angle — coupled with an increasing need to create and adapt promising new drug-discovery technologies — has led to some relatively unconventional recruitment at Merck. For instance, the company brought in an engineer from Oak Ridge National Laboratory in Tennessee to help

miniaturize the company's drug-screening operation. And a Merck subsidiary, Seattle-based Rosetta Inpharmatics, hired several astrophysicists to beef up its bioinformatics tools.

To hire people with backgrounds oriented towards problem-solving rather than specific technology proficiency, Merck is looking increasingly for scientists with dual or mixed training — someone who did a first degree in chemical engineering and a PhD in biology, for example. "These people tend to work out extremely well," says Ford-Hutchinson.

Cellzome, a biotech company based in Heidelberg, Germany, is also now recruiting from an array of scientific disciplines. It is searching for people who can use its core technology — functional proteomics hardware and software — and any other available scientific techniques to ask questions about how drugs interact with protein complexes, says Walter Blackstock, Cellzome's vice-president of technology. "We need people to design cool experiments," he says.

BUILT-IN OBSOLESCENCE

Genetic engineering provides a good example of how the demand for technical know-how, rather than the broader ability to ask scientific questions, fades as technology changes. People who had the skills to make knockout mice were once much in demand, says Richard Scheller, senior vice-president for research at Genentech, a San Francisco-based biotech company. Now off-the-shelf kits have made that sort of molecular-biology specialist less important.

Having the ability to add a scientific angle to that technical expertise replaces some of the lost value. For instance, people who can make transgenic mice that accurately reflect different human diseases are very much in demand, Scheller says. He agrees that general problem-solving skills are sought by the drug-development industry, but notes that each company has its own specific requirements as well.

Different companies could have widely varying needs, he adds, depending on the therapeutic areas or disease-fighting mechanisms they emphasize. For example, one of Genentech's main strategies is antibody therapeutics for cancer, so the company is always looking for people with a research background in oncology or immunology.

But like many biotechs, Genentech wants to expand



into new kinds of treatment. Although the company's existing therapeutics are protein-based, its researchers are now looking for ways to design small molecules that mimic antibody binding. This approach opens the door for people with additional skills — including the two structural biologists hired recently to assist in this rational approach to drug design. And if the structural biologists generate drug targets, the company will require medicinal chemists to make the small molecules that mimic the drugs.

Scheller says that medicinal chemists can often find work straight after finishing their PhD. If the demand for such scientists increases further among the biotech firms, then large pharmaceutical companies, who are already finding it hard to recruit enough of these skills themselves, face a serious skills shortage.

"There's a big concern about the numbers of medicinal chemists," says Martin. The ones recruited to Abbott usually need to learn some biology so they can

The search for fresh therapeutics is opening the door for a range of disciplines.

Instrumental in instrumentation

As the efforts of pharmaceutical companies to develop new drugs become increasingly reliant on instrument manufacturers, the recruitment needs for the two sectors are starting to converge. And both are competing for talent from some of the same pools — especially bioinformaticians, chemists and cell biologists. In addition, development, sales and support people all tend to

have more scientific training than in the past.

John Cerio, vice-president of human resources at instrumentation company PerkinElmer Life Sciences in Boston, says that the demand for people who have the ability to develop new instruments and assays is now "robust", as the company aims to speed new products to market. Speed is perhaps the biggest difference

between working in each sector.

In drug discovery, development is measured in years, but in instrumentation, the deadlines for new products are viewed in months or even weeks. People who want to work in instrumentation must have "a strong sense of urgency", Cerio says.

Ian Taylor, group product manager at PerkinElmer, notes that skills in sales are also

leaning more towards the scientific, because salespeople now serve more of a "consulting" role. They have to know how instruments can be used to answer specific scientific questions and also be able to help customers set them up and test them.

"The people who are the most successful have PhDs, can talk the science and understand the project," Taylor says. **P.S.**

communicate with the structural people.

The shortage in certain skills tends to show regional variation. Mads Krogsgaard Thomsen, executive vice-president and chief scientific officer for Novo Nordisk, a health-care company based in Bagsværd, Denmark, points to a survey conducted last September by the Association of the Danish Pharmaceutical Industry. This showed that within Medicon Valley — a biotech hotspot that straddles Sweden and Denmark (see *Naturejobs* 10–11, 21 June 2001) — there would be a shortage of several hundred PhD-level engineers within the next five years.

He also suspects that there are fewer people conversant with both biology and intellectual-property issues in Europe than in the United States, which has a longer history of university-supported technology transfer. “In the United States this may be more embedded, but in most European countries such thinking is still in its infancy,” Krogsgaard Thomsen says.

Regional differences also apply when looking for people, in particular physicians, who can work in the clinical-development phase. The United States and Europe have some experienced people who are very much in demand, whereas in Japan these specialists are “impossible” to find, he comments.

GROWING TALENT

But people with certain skills, such as *in vivo* pharmacologists, who work with animals, are difficult to find anywhere. So Novo Nordisk, like many other companies, is growing its own, using studentships. The company sponsors PhD programmes in academic centres and gives the students part-time employment during their studies. “We end up recruiting three out of four,” Krogsgaard Thomsen says.

Franz Hefti, senior vice-president for neuroscience research at Merck Research Laboratories in San Diego, agrees that recruiting good *in vivo* pharmacologists is difficult. “People who can work with animals are very, very scarce,” he says. Merck also funds studentships and research fellowships that focus on *in vivo* pharmacology.

The biggest skill that companies need is perhaps the most general — and so the most difficult to teach: an understanding of the drug-discovery process. “Drug discovery is quite different from what you do in a research lab,” Hefti says. “Surprisingly few people know that.”

Franz Hefti finds it hard to recruit *in vivo* pharmacologists.



Opportunities in transformations

When technology-based companies reinvent themselves as drug-discovery companies, they often find they have many of the skills established drug-discovery companies desire — but lack many of the core competencies. So to complete the transformation, they seek to fill that gap.

For example, Structural GenomiX, a San Diego-based biotech company, already had a technology platform to discern protein structure and the scientific skills to use it. And it bought San Diego-based Prospect Genomics for bioinformatics punch.

Bioinformaticians and structural biologists are in demand at established drug-discovery firms. But Structural GenomiX is now finding itself looking for several kinds of chemists, the group of scientists that most drug-discovery companies build themselves around.

Similarly, Celera, based in Rockville, Maryland, which used to focus on genomic information, will have to find traditional drug-discovery people to augment its staff of sequencers and bioinformaticians as it turns its attention to the search for therapeutics.

P.S.

One of the difficulties is that those skills change as a drug progresses through the pipeline. And in biotech firms, which typically have fewer employees and specialists than drug companies, newer employees need to learn other aspects, such as regulatory requirements, says Tomi Sawyer, a vice-president at ARIAD Pharmaceuticals in Cambridge, Massachusetts.

Sawyer, who has previously taught courses focused on drug discovery at several universities, notes that there is a world of difference between labs in academia, which are oriented towards basic research, and those in industry. Moving from one to the other is like “playing backyard football one day, then playing with the professionals at the Super Bowl the next”.

But drug-discovery professionals can thrive by practising existing skills and learning new ones. ■

Paul Smaglik is editor of *Naturejobs*.

Adam Smith is editor of *Nature Reviews Drug Discovery*.



MERCK

Life beyond the walls

Many large drug companies are now taking advantage of the specialist skills offered by smaller biotech firms in their search for new drugs. This trend for outsourcing elements of research is forcing a careful evaluation of licensing policy, says Adam Smith.

Over the past couple of decades, in parallel with the 'biotech explosion', large pharmaceutical companies have become more willing to look outside their own walls for ideas and leads to fuel drug development (see *Nature* **414**, 482–483; 2001). It makes sense for them to benefit from the specialist technology or knowledge bases offered by smaller companies. And the smaller biotech firms can gain from the help offered by their larger partners when it comes to turning the therapeutic opportunities they have identified into commercial products.

At one time the shift towards outsourcing discovery efforts looked like threatening the existence of the in-house teams at big companies. "Five or six years ago, prestigious leaders in the pharmaceutical industry were talking about the 'paradigm shift', in which pharmaceutical companies would be seen as drug-development warehouses and biotechs as the

creative engines," says Tomi Sawyer, a vice-president at ARIAD Pharmaceuticals in Cambridge, Massachusetts.

But what has actually emerged is a more symbiotic relationship. According to Franz Hefti, senior vice-president for neuroscience research at Merck Research Laboratories in San Diego: "Biotechs now compete with the internal discovery programmes of large pharmaceutical firms."



Tomi Sawyer: smaller companies are attempting to develop their own drugs.

One of the weapons used by big drug companies to maintain their competitive edge over biotech firms is the PhDs they employ to scout for licensing opportunities. Until the mid-1990s, licensing departments tended to react to biotech's advances. Now they've become much more proactive, looking at everything from basic science to drugs ready for development. The Danish health-care company Novo Nordisk has underlined the importance of licensing basic science and discovery research by separating its

scientific licensing department from the more established commercial licensing of drug-development opportunities.

To benefit from this competition, pharmaceutical companies need to be sure that they understand what is being offered to them, opting to license only the very best ideas. "The critical factor for this approach to be successful is that, without an in-house knowledge base, the likelihood of harvesting the fruits of scientific licensing is very small," says Mads Krogsgaard Thomsen, Novo Nordisk's executive vice-president and chief scientific officer. Maintaining a critical mass of discovery scientists within the company acts as a "sanity check" on the licensing decisions, according to Krogsgaard Thomsen. "You can only outsource up to a level that you can still control," he says.

The prediction of a paradigm shift may also have underestimated biotech firms' desire to be more than just 'creative engines'. "A lot of small companies are

taking the high-risk, high-profit strategy of trying to develop their own drugs," says Sawyer. Summing up the sense that biotech discovery efforts will certainly complement, but never replace, discovery research in big pharmaceutical operations, Sawyer says: "I don't know if anything has changed in 10 years, in terms of the skill set that is needed to join a large pharmaceutical company."



In-house knowledge is vital, according to Mads Krogsgaard Thomsen.

Some larger companies, such as GlaxoSmithKline and Novartis, are busy trying to capture the entrepreneurial spirit that smallness fosters by redesigning their discovery teams worldwide into a collection of 'biotech-like' groupings.

Another driver for such change might be the perception of a more flexible working environment within biotech, and the desire to keep employees in the pharmaceutical industry happy and motivated and in "less of a coal-miner situation", as Hefti puts it. ■

Adam Smith is editor of *Nature Reviews Drug Discovery*.